Optimizing Black-box Metrics with Adaptive Surrogates

Qijia Jiang, Olaoluwa Adigun, Harikrishna Narasimhan, Mahdi Milani Fard, Maya Gupta

Appendix

Notations. We use $[K]$ to denote $\{1, \ldots, K\}$. We use $\|\cdot\|$ to denote the L_2-norm. Unless specified otherwise, all smoothness and Lispchitz definitions are with respect to the L_2-norm.

A. Proofs for Theorems and Lemmas

A.1. Proof of Observation 1

Proof. To see that the vector $[u_1, \cdots, u_K]$ belongs to a convex set, since by assumption $\{\ell_i\}_{i=1}^K$ are convex functions, therefore the set of constraints $\ell_i(\theta) \leq u_i$ defines a convex set in $[\theta, u_1, \cdots, u_K]$ as intersection of sublevel sets of convex functions are convex. □

A.2. Proof of Lemma 1

Lemma 1 (Restated). *Let $\mathbf{u}^+$ be the exact projection of $\tilde{\mathbf{u}}^{t+1} \in \mathbb{R}_+^K$ onto $\mathcal{U}$. For any solution $\mathbf{u}^{t+1}$ to (4), we have $\mathbf{u}^{t+1} \in \mathcal{U}$, $\mathbf{u}^{t+1} \leq \mathbf{u}^+$, and for a monotonic ψ, $\psi(\mathbf{u}^{t+1}) \leq \psi(\mathbf{u}^+)$.*

We first show how one can compute an exact projection onto $\mathcal{U}$, and show that the projection described in Lemma 1 implements this approximately.

Lemma 5 (Exact projection). *The projection $\mathbf{u}^+$ of $\tilde{\mathbf{u}}^{t+1} \in \mathbb{R}_+^K$ onto $\mathcal{U}$ is given by:*

(i) $\theta^{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{2} \| \big(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^{t+1}\big)_+ \|^2$; $\quad \mathbf{u}^{t+1} = \boldsymbol{\ell}(\theta^{t+1})$

(ii) $u_k^+ = \max\{\tilde{u}_k, u_k^{t+1}\}$, $\forall k \in [K]$,

where $(z)_+ = \max\{0, z\}$, applied element-wise.

Proof. It is easy to see that step (i) is a convex problem because each ℓ_k is convex in θ, and both $(\cdot)_+$ and $\|\cdot\|^2$ are convex and monotonic in their arguments, making the composition $\| \big(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}\big)_+ \|^2$ also convex in θ.

To perform the projection, since step (i) above is a convex problem, the optimality condition gives

$$\sum_{i=1}^K (\ell_i(\theta^+) - \tilde{u}_i)_+ \cdot \mathbb{1}\{\ell_i(\theta^+) - \tilde{u}_i > 0\} \cdot \nabla_\theta \ell_i(\theta^+) = \mathbf{0}_d$$

which is the same as

$$\sum_{i=1}^K (u_i^+ - \tilde{u}_i) \cdot \nabla_\theta \ell_i(\theta^+) = \mathbf{0}_d \tag{5}$$

by the second step of the procedure. We shall use (5) to show that u_i^+ is the projection in the $\mathcal{U}$-space.

The projection in the $\mathcal{U}$-space can equivalently be written as the following convex problem

$$\begin{aligned} \underset{u_1, \cdots, u_K, \theta}{\text{minimize}} \quad & \frac{1}{2} \sum_{i=1}^K (u_i - \tilde{u}_i)^2 \\ \text{subject to} \quad & u_i - \ell_i(\theta) \geq 0 \quad \forall i \in [K]\,. \end{aligned}$$

Introducing the dual variable $\lambda \in \mathbb{R}^K$ and the KKT condition of the problem becomes

$$\sum_{i=1}^K (u_i - \tilde{u}_i) - \sum_{i=1}^K \lambda_i = 0 \qquad \sum_{i=1}^K \lambda_i \cdot \nabla_\theta \ell_i(\theta) = 0$$

$$u_i - \ell_i(\theta) \geq 0 \qquad \lambda_i \geq 0 \qquad \lambda_i(u_i - \ell_i(\theta)) = 0 \quad \forall i \in [K]$$

if $(u_1, \cdots, u_K, \theta)$ and λ are optimal.

Taking $\lambda_i = u_i^+ - \tilde{u}_i$ and $\theta = \theta^+$ with $u_i = u_i^+$, one can easily verify using (5) that all the conditions hold. Since the optimization problem satisfies Slater's constraint qualification and therefore we can conclude that the primal optimal solution is $\mathbf{u}^+$, as defined in the lemma statement. □

We go on to prove Lemma 1.

Proof of Lemma 1. Because $\mathbf{u}^{t+1}$ is the surrogate loss at θ^{t+1}, it clearly lies in $\mathcal{L}$ and hence in the superset $\mathcal{U} \supseteq \mathcal{L}$. Next, notice that the over-constrained projection $\mathbf{u}^{t+1}$ in Lemma 1 is the same as step (i) in the exact projection in Lemma 1, with step (ii) giving us that the exact projection $u_k^+ = \max\{\tilde{u}_k,\, u_k^{t+1}\},\ \forall k \in [K]$. It follows that: $u_k^{t+1} \leq u_k^+, \forall k \in [K]$. So for a monotonic ψ, we have $\psi(\mathbf{u}^{t+1}) \leq \psi(\mathbf{u}^+)$. □

A.3. Proof of Theorem 2

Theorem 2 (Restated)**.** *Let $M(\theta) = \psi(\boldsymbol{\ell}(\theta)) + \epsilon(\theta)$, for a ψ that is monotonic, β-smooth and L-Lipschitz, and the worst-case slack $\max_{\theta \in \mathbb{R}^d} |\epsilon(\theta)|$ is the minimum among all such decompositions of M.*

Suppose each ℓ_k is γ-smooth and Φ-Lipschitz in θ with $\|\boldsymbol{\ell}(\theta)\| \leq G, \forall \theta$. Suppose the gradient estimates $\hat{\mathbf{g}}^t$ satisfy $\mathbf{E}\left[\|\hat{\mathbf{g}}^t - \nabla\psi(\boldsymbol{\ell}(\theta^t))\|^2\right] \leq \kappa_\epsilon,\ \forall t \in [T]$ and the projection step satisfies $\|(\boldsymbol{\ell}(\theta^{t+1}) - \tilde{\mathbf{u}}^t)_+\|^2 \leq \min_{\theta \in \mathbb{R}^d} \|(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^t)_+\|^2 + \mathcal{O}(\frac{1}{\beta^2 T}),\ \forall t \in [T]$. Set stepsize $\eta = \frac{1}{\beta^2}$.

Then Algorithm 1 converges to an approximate stationary point of $\psi(\boldsymbol{\ell}(\cdot))$:

$$\min_{1 \leq t \leq T} \mathbf{E}\left[\|\nabla\psi(\boldsymbol{\ell}(\theta^t))\|^2\right] \leq C\left(\frac{\beta}{\sqrt{T}} + \sqrt{\kappa_\epsilon} + \sqrt{L}\kappa_\epsilon^{1/4}\right),$$

where the expectation is over the randomness in the gradient estimates, and $C = \mathcal{O}\big(KL\big(\gamma\big(G + \frac{L}{\beta^2}\big) + \Phi^2\big)\big)$.

While the above theorem prescribes a specific learning rate η for the projected gradient descent, in our experiments, we tune η using a held-out validation set.

The proof proceeds in two parts. In Section A.3.1, we first show that the algorithm converges to an approximate stationary point of ψ over $\mathcal{U}$. In Section A.3.2, we then translate this a guarantee in θ, i.e. we show that the algorithm converges to an approximate stationary point of $\psi(\boldsymbol{\ell}(\cdot))$ over θ.

A.3.1. Convergence in $\mathcal{U}$-space

Lemma 6. *Define the gradient mapping at $\mathbf{u} \in \mathcal{U}$ for a vector $g \in \mathbb{R}^K$ as $P(\mathbf{u},\, g) := \frac{1}{\eta}(\mathbf{u} - \Pi_{\mathcal{U}}(\mathbf{u} - \eta \cdot g))$, where $\Pi_{\mathcal{U}}(z)$ denotes the projection of z onto $\mathcal{U}$. Then under the assumptions of Theorem 2,*

$$\min_{1 \leq t \leq T} \mathbf{E}\left[\|P(\mathbf{u}^t, \nabla\psi(\mathbf{u}^t))\|^2\right] \leq \mathcal{O}\left(\frac{\beta^2}{T} + \kappa_\epsilon + L\sqrt{\kappa_\epsilon}\right).$$

Before we prove this result, we will find it useful to state the following lemma.

Lemma 7 (Properties of inexact projection)**.** *Fix $\mathbf{u} \in \mathcal{U}$ where $\mathcal{U}$ is a convex set and arbitrary vectors $g_1, g_2 \in \mathbb{R}^K$. Let*

$$\theta_1^+ \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{2}\|(\boldsymbol{\ell}(\theta) - (\mathbf{u} - \eta g_1))_+\|^2 \quad \text{and} \quad \theta_2^+ \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{2}\|(\boldsymbol{\ell}(\theta) - (\mathbf{u} - \eta g_2))_+\|^2,$$

and let $\mathbf{u}_1^+ = \max\{\boldsymbol{\ell}(\theta_1^+),\, \mathbf{u} - \eta g_1\}$ and $\mathbf{u}_2^+ = \max\{\boldsymbol{\ell}(\theta_2^+),\, \mathbf{u} - \eta g_2\}$. Define the gradient mapping $P(\mathbf{u},\, g_1) := \frac{1}{\eta}(\mathbf{u} - \mathbf{u}_1^+)$ and $P(\mathbf{u},\, g_2) := \frac{1}{\eta}(\mathbf{u} - \mathbf{u}_2^+)$. Denote $\tilde{\theta}_1^+, \tilde{\theta}_2^+$ as approximate minimizers such that

$$\frac{1}{2}\|(\boldsymbol{\ell}(\tilde{\theta}_1^+) - (\mathbf{u} - \eta g_1))_+\|^2 \leq \frac{1}{2}\|(\boldsymbol{\ell}(\theta_1^+) - (\mathbf{u} - \eta g_1))_+\|^2 + \alpha \tag{6}$$

and

$$\frac{1}{2}\|(\boldsymbol{\ell}(\tilde{\theta}_2^+) - (\mathbf{u} - \eta g_2))_+\|^2 \leq \frac{1}{2}\|(\boldsymbol{\ell}(\theta_2^+) - (\mathbf{u} - \eta g_2))_+\|^2 + \alpha, \tag{7}$$

and let $\tilde{\mathbf{u}}_1^+ = \max\{\boldsymbol{\ell}(\tilde{\theta}_1^+), \mathbf{u} - \eta g_1\}$ *and* $\tilde{\mathbf{u}}_2^+ = \max\{\boldsymbol{\ell}(\tilde{\theta}_2^+), \mathbf{u} - \eta g_2\}$. *Define the corresponding gradient mapping* $\tilde{P}(\mathbf{u}, g_1) := \frac{1}{\eta}(\mathbf{u} - \tilde{\mathbf{u}}_1^+)$ *and* $\tilde{P}(\mathbf{u}, g_2) := \frac{1}{\eta}(\mathbf{u} - \tilde{\mathbf{u}}_2^+)$. *Then the following holds:*

1. $\|\tilde{P}(\mathbf{u}, g_1) - P(\mathbf{u}, g_1)\| \leq \frac{\sqrt{2\alpha}}{\eta}$.
2. $\langle g_1, \tilde{P}(\mathbf{u}, g_1)\rangle \geq \frac{3}{4}\|\tilde{P}(\mathbf{u}, g_1)\|^2 - \frac{2\alpha}{\eta^2}$.
3. $\|\tilde{P}(\mathbf{u}, g_1)\| \leq \|g_1\| + \frac{\sqrt{2\alpha}}{\eta}$.
4. $\|P(\mathbf{u}, g_1) - P(\mathbf{u}, g_2)\| \leq \|g_1 - g_2\|$.
5. $\|\tilde{P}(\mathbf{u}, g_1) - \tilde{P}(\mathbf{u}, g_2)\| \leq \|g_1 - g_2\| + 2\frac{\sqrt{2\alpha}}{\eta}$.

Proof. We have:

$$\begin{aligned}
\frac{1}{2}\|\tilde{\mathbf{u}}_1^+ - (\mathbf{u} - \eta g_1)\|^2 &= \frac{1}{2}\|\max\{\mathbf{u} - \eta g_1, \boldsymbol{\ell}(\tilde{\theta}_1^+)\} - (\mathbf{u} - \eta g_1)\|^2 \\
&= \frac{1}{2}\|(\boldsymbol{\ell}(\tilde{\theta}_1^+) - (\mathbf{u} - \eta g_1))_+\|^2 \\
&\leq \frac{1}{2}\|(\boldsymbol{\ell}(\theta_1^+) - (\mathbf{u} - \eta g_1))_+\|^2 + \alpha \qquad \text{(Assumption (6))} \\
&= \frac{1}{2}\|\max\{\mathbf{u} - \eta g_1, \boldsymbol{\ell}(\theta_1^+)\} - (\mathbf{u} - \eta g_1)\|^2 + \alpha \\
&= \frac{1}{2}\|\mathbf{u}_1^+ - (\mathbf{u} - \eta g_1)\|^2 + \alpha,
\end{aligned}$$

which implies that

$$g_1^\top \tilde{\mathbf{u}}_1^+ + \frac{1}{2\eta}\|\tilde{\mathbf{u}}_1^+ - \mathbf{u}\|^2 - g_1^\top \mathbf{u}_1^+ - \frac{1}{2\eta}\|\mathbf{u}_1^+ - \mathbf{u}\|^2 \leq \frac{\alpha}{\eta}\,. \tag{8}$$

Part (1) now follows from

$$\begin{aligned}
\|\tilde{P}(\mathbf{u}, g_1) - P(\mathbf{u}, g_1)\| &= \frac{1}{\eta}\|\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+\| \\
&\leq \frac{\sqrt{2\eta}}{\eta}\sqrt{F_{g_1}(\tilde{\mathbf{u}}_1^+) - F_{g_1}(\mathbf{u}_1^+) - \nabla F_{g_1}(\mathbf{u}_1^+)^\top(\tilde{\mathbf{u}}_1^+ - \mathbf{u}_1^+)} \\
&\leq \frac{\sqrt{2\eta}}{\eta}\sqrt{\frac{\alpha}{\eta}} \leq \frac{\sqrt{2\alpha}}{\eta}\,,
\end{aligned}$$

where we used $\frac{1}{\eta}$-strong convexity of the objective $F_{g_1}(\mathbf{z}) := g_1^\top \mathbf{z} + \frac{1}{2\eta}\|\mathbf{z} - \mathbf{u}\|^2$ for $\mathbf{z}, \mathbf{u} \in \mathcal{U}$ and the fact that $\mathbf{u}_1^+$ is the exact minimizer over the convex set $\mathcal{U}$, implying $\nabla F_{g_1}(\mathbf{u}_1^+)^\top(\mathbf{z} - \mathbf{u}_1^+) \geq 0\ \forall \mathbf{z} \in \mathcal{U}$.
For part (4), since $\mathbf{u}_1^+$ and $\mathbf{u}_2^+$ are optimal points of function $F_{g_1}(\cdot)$ and $F_{g_2}(\cdot)$ over convex set $\mathcal{U}$ respectively, from optimality condition we have

$$\Big(g_1 + \frac{1}{\eta}(\mathbf{u}_1^+ - \mathbf{u})\Big)^\top(\mathbf{z} - \mathbf{u}_1^+) \geq 0 \quad \text{and} \quad \Big(g_2 + \frac{1}{\eta}(\mathbf{u}_2^+ - \mathbf{u})\Big)^\top(\mathbf{z} - \mathbf{u}_2^+) \geq 0 \quad \text{for all } \mathbf{z} \in \mathcal{U}\,. \tag{9}$$

Setting $\mathbf{z} = \mathbf{u}_2^+$ in the first and $\mathbf{z} = \mathbf{u}_1^+$ in the second equation and summing up we have

$$(g_1 - g_2)^\top(\mathbf{u}_2^+ - \mathbf{u}_1^+) \geq \frac{1}{\eta}\|\mathbf{u}_2^+ - \mathbf{u}_1^+\|^2\,.$$

Therefore using Cauchy-Schwarz

$$\|P(\mathbf{u}, g_1) - P(\mathbf{u}, g_2)\| = \frac{1}{\eta}\|\mathbf{u}_2^+ - \mathbf{u}_1^+\| \leq \|g_1 - g_2\|\,.$$

Part (5) now follows immediately from part (1) and (4) by

$$\begin{aligned}\|\tilde{P}(\mathbf{u}, g_1) - \tilde{P}(\mathbf{u}, g_2)\| &\leq \|P(\mathbf{u}, g_1) - P(\mathbf{u}, g_2)\| + \|\tilde{P}(\mathbf{u}, g_1) - P(\mathbf{u}, g_1) + P(\mathbf{u}, g_2) - \tilde{P}(\mathbf{u}, g_2)\| \\ &\leq \|g_1 - g_2\| + 2\|\tilde{P}(\mathbf{u}, g_1) - P(\mathbf{u}, g_1)\| \\ &\leq \|g_1 - g_2\| + \frac{2\sqrt{2\alpha}}{\eta}\,.\end{aligned}$$

To see part (2), we plug in $\mathbf{z} = \mathbf{u}$ in the first equation of display (9), giving $g_1^\top(\mathbf{u} - \mathbf{u}_1^+) \geq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2$. Moreover from equation (8) we know

$$g_1^\top(\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+) \geq -\frac{\alpha}{\eta} + \frac{1}{2\eta}\|\tilde{\mathbf{u}}_1^+ - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{u}_1^+ - \mathbf{u}\|^2\,.$$

Consequently,

$$g_1^\top(\mathbf{u} - \tilde{\mathbf{u}}_1^+) = g_1^\top(\mathbf{u} - \mathbf{u}_1^+) + g_1^\top(\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+) \geq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2 - \frac{\alpha}{\eta} + \frac{1}{2\eta}\|\tilde{\mathbf{u}}_1^+ - \mathbf{u}\|^2 - \frac{1}{2\eta}\|\mathbf{u}_1^+ - \mathbf{u}\|^2\,.$$

Now to relate $\|\mathbf{u} - \mathbf{u}_1^+\|$ to $\|\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+\|$, we have

$$\begin{aligned}\frac{1}{2\eta}\|\mathbf{u} - \tilde{\mathbf{u}}_1^+\|^2 &\leq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2 + \frac{1}{\eta}\|\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+\|^2 \\ &\leq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2 + 2[F_{g_1}(\tilde{\mathbf{u}}_1^+) - F_{g_1}(\mathbf{u}_1^+) - \nabla F_{g_1}(\mathbf{u}_1^+)^\top(\tilde{\mathbf{u}}_1^+ - \mathbf{u}_1^+)] \\ &\leq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2 + \frac{2\alpha}{\eta}\,.\end{aligned}$$

Putting things together $g_1^\top \tilde{P}(\mathbf{u}, g_1) = \frac{1}{\eta} g_1^\top(\mathbf{u} - \tilde{\mathbf{u}}_1^+) \geq \frac{3}{4}\|\tilde{P}(\mathbf{u}, g_1)\|^2 - \frac{2\alpha}{\eta^2}$, as claimed.
Finally, for part (3) since $\|g_1\| \cdot \|\mathbf{u} - \mathbf{u}_1^+\| \geq g_1^\top(\mathbf{u} - \mathbf{u}_1^+) \geq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\|^2$ and using part (2),

$$\begin{aligned}\|\tilde{P}(\mathbf{u},\, g_1)\| = \frac{1}{\eta}\|\mathbf{u} - \tilde{\mathbf{u}}_1^+\| &\leq \frac{1}{\eta}\|\mathbf{u} - \mathbf{u}_1^+\| + \frac{1}{\eta}\|\mathbf{u}_1^+ - \tilde{\mathbf{u}}_1^+\| \\ &\leq \|g_1\| + \frac{\sqrt{2\alpha}}{\eta}\,,\end{aligned}$$

where we used part (1) for the last step. This concludes the proof of the lemma. □

Equipped with the above results, we move on to prove Lemma 6, i.e. to show that the algorithm converges to an approximate stationary point of ψ over $\mathcal{U}$.

Proof of Lemma 6. We will assume that the gradient estimates $\hat{\mathbf{g}}^t$ satsify $\mathbf{E}\left[\|\hat{\mathbf{g}}^t - \nabla\psi(\boldsymbol{\ell}(\theta^t))\|^2\right] \leq \kappa_\epsilon,\ \forall t \in [T]$ and the projection step satisfies $\frac{1}{2}\|(\boldsymbol{\ell}(\theta^{t+1}) - \tilde{\mathbf{u}}^t)_+\|^2 \leq \min_{\theta\in\mathbb{R}^d} \frac{1}{2}\|(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^t)_+\|^2 + \alpha,\ \forall t \in [T]$.

Let $\mathbf{u}^{t+1} = \boldsymbol{\ell}(\theta^{t+1})$ and $\tilde{\mathbf{u}}^{t+1} = \max\{\mathbf{u}^{t+1}, \mathbf{u}^t - \eta\hat{\mathbf{g}}^t\}$ be the next iterate had we executed step (ii) of the projection given Lemma 1. Define $\delta^t := \hat{\mathbf{g}}^t - \nabla\psi(\mathbf{u}^t)$. For any $g \in \mathbb{R}^K$, let the gradient mapping $P(\mathbf{u}, g)$ and approximate gradient

mapping $\tilde{P}(\mathbf{u}, g)$ be defined as in Lemma 7. Note that $\tilde{\mathbf{u}}^{t+1} = \mathbf{u}^t - \eta \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)$.

$$
\begin{aligned}
\psi(\mathbf{u}^{t+1}) &\leq \psi(\tilde{\mathbf{u}}^{t+1}) \quad \text{(from monotonicity of } \psi\text{)} \\
&\leq \psi(\mathbf{u}^t) - \eta \langle \nabla\psi(\mathbf{u}^t),\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \frac{\beta^2}{2}\eta^2 \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 \quad \text{(using smoothness of } \psi\text{)} \\
&= \psi(\mathbf{u}^t) - \eta \langle \hat{\mathbf{g}}^t,\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \eta \langle \hat{\mathbf{g}}^t - \nabla\psi(\mathbf{u}^t),\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \frac{\beta^2}{2}\eta^2 \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 \\
&= \psi(\mathbf{u}^t) - \eta \langle \hat{\mathbf{g}}^t,\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \eta \langle \delta^t,\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \frac{\beta^2}{2}\eta^2 \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 \\
&\leq \psi(\mathbf{u}^t) - \left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 + \eta \langle \delta^t,\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\rangle + \frac{2\alpha}{\eta} \quad \text{(from Lemma 7, statement 2)} \\
&= \psi(\mathbf{u}^t) - \left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 + \eta \langle \delta^t,\, \tilde{P}(\mathbf{u}^t, \nabla\psi(\mathbf{u}^t))\rangle + \eta \langle \delta^t,\, \tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t) - \tilde{P}(\mathbf{u}^t, \nabla\psi(\mathbf{u}^t))\rangle + \frac{2\alpha}{\eta} \\
&\leq \psi(\mathbf{u}^t) - \left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 + \eta \langle \delta^t,\, \tilde{P}(\mathbf{u}^t, \nabla\psi(\mathbf{u}^t))\rangle + \eta\|\delta^t\|^2 + 2\sqrt{2\alpha}\,\|\delta^t\| + \frac{2\alpha}{\eta} \\
&\leq \psi(\mathbf{u}^t) - \left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 + \eta\|\delta^t\|\left(\|\nabla\psi(\mathbf{u}^t)\| + \frac{\sqrt{2\alpha}}{\eta}\right) + \eta\|\delta^t\|^2 + 2\sqrt{2\alpha}\,\|\delta^t\| + \frac{2\alpha}{\eta} \\
&\leq \psi(\mathbf{u}^t) - \left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 + (\eta L + \sqrt{2\alpha})\|\delta^t\| + \eta\|\delta^t\|^2 + 2\sqrt{2\alpha}\,\|\delta^t\| + \frac{2\alpha}{\eta},
\end{aligned}
$$

where the third-last inequality uses Lemma 7, statement 5 together with Cauchy-Schwarz and the second-last inequality uses Lemma 7, statement 3, and the fact that ψ is L-Lipschitz. Summing up over $t = 1, \ldots, T$,

$$
\left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \sum_{t=1}^{T} \|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2 \leq \psi(\mathbf{u}^1) - \psi(\mathbf{u}^{t+1}) + \sum_{t=1}^{T} \left((\eta L + 3\sqrt{2\alpha})\|\delta^t\| + \eta\|\delta^t\|^2 + \frac{2\alpha}{\eta}\right).
$$

Taking expectations on both sides and using the assumption $0 \leq \psi(\mathbf{u}) \leq 1\ \forall \mathbf{u} \in \mathcal{U}$,

$$
\begin{aligned}
\left(\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2\right) \sum_{t=1}^{T} \mathbf{E}\left[\|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] &\leq 1 + \sum_{t=1}^{T} \left((\eta L + 3\sqrt{2\alpha})\mathbf{E}\left[\|\delta^t\|\right] + \eta\mathbf{E}\left[\|\delta^t\|^2\right] + \frac{2\alpha}{\eta}\right) \\
&\leq 1 + \sum_{t=1}^{T} \left((\eta L + 3\sqrt{2\alpha})\sqrt{\mathbf{E}\left[\|\delta^t\|^2\right]} + \eta\mathbf{E}\left[\|\delta^t\|^2\right] + \frac{2\alpha}{\eta}\right) \\
&\leq 1 + T\left((\eta L + 3\sqrt{2\alpha})\sqrt{\kappa_\epsilon} + \eta\kappa_\epsilon + \frac{2\alpha}{\eta}\right),
\end{aligned}
$$

where we used the assumption on the gradient estimate error $\mathbf{E}\left[\|\delta^t\|^2\right]$ in the last step. Rearranging we have

$$
\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[\|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] \leq \frac{1/T + (\eta L + 3\sqrt{2\alpha})\sqrt{\kappa_\epsilon} + \eta\kappa_\epsilon + \frac{2\alpha}{\eta}}{\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2}.
$$

Using Lemma 7, statement 1,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[\|P(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] &\leq \frac{2}{T}\sum_{t=1}^{T} \mathbf{E}\left[\|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] + \frac{2}{T}\sum_{t=1}^{T} \mathbf{E}\left[\|\tilde{P}(\mathbf{u}^t, \hat{\mathbf{g}}^t) - P(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] \\
&\leq \frac{2/T + 2(\eta L + 3\sqrt{2\alpha})\sqrt{\kappa_\epsilon} + 2\eta\kappa_\epsilon + \frac{4\alpha}{\eta}}{\frac{3}{4}\eta - \frac{\beta^2}{2}\eta^2} + \frac{4\alpha}{\eta^2}.
\end{aligned}
$$

Setting stepsize $\eta = \frac{1}{\beta^2}$:

$$
\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[\|P(\mathbf{u}^t, \hat{\mathbf{g}}^t)\|^2\right] \leq \frac{8\beta^2}{T} + 8L\sqrt{\kappa_\epsilon} + 8\kappa_\epsilon + 24\beta^2\sqrt{2\alpha\kappa_\epsilon} + 20\alpha\beta^4\,.
$$

We can now bound the average gradient map norm across iterations:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\left[\|P(\mathbf{u}^t,\nabla\psi(\mathbf{u}^t))\|^2\right] &\leq \frac{2}{T}\sum_{t=1}^{T}\mathbf{E}\left[\|P(\mathbf{u}^t,\hat{\mathbf{g}}^t)\|^2\right] + \frac{2}{T}\sum_{t=1}^{T}\mathbf{E}\left[\|P(\mathbf{u}^t,\nabla\psi(\mathbf{u}^t)) - P(\mathbf{u}^t,\hat{\mathbf{g}}^t)\|^2\right]\\ &\leq \frac{2}{T}\sum_{t=1}^{T}\mathbf{E}\left[\|P(\mathbf{u}^t,\hat{\mathbf{g}}^t)\|^2\right] + \frac{2}{T}\sum_{t=1}^{T}\mathbf{E}\left[\|\nabla\psi(\mathbf{u}^t) - \hat{\mathbf{g}}^t\|^2\right]\\ &\leq \frac{16\beta^2}{T} + 16L\sqrt{\kappa_\epsilon} + 16\kappa_\epsilon + 48\beta^2\sqrt{2\alpha\kappa_\epsilon} + 40\alpha\beta^4 + 2\kappa_\epsilon\end{aligned}$$

where we used Lemma 7, statement 4 for the second inequality and the assumption on the gradient estimation error for the last inequality. Thus:

$$\min_{1\leq t\leq T}\mathbf{E}\left[\|P(\mathbf{u}^t,\nabla\psi(\mathbf{u}^t))\|^2\right] \leq \frac{16\beta^2}{T} + 16L\sqrt{\kappa_\epsilon} + 18\kappa_\epsilon + 48\beta^2\sqrt{2\alpha\kappa_\epsilon} + 40\alpha\beta^4.$$

Now picking $\alpha = \frac{1}{\beta^2 T}$ completes the proof. □

A.3.2. Convergence in θ-space

We are now ready to prove Theorem 2. We translate the near-stationarity result in Lemma from $\mathbf{u}$-space to θ-space.

Proof of Theorem 2. For a given T, let $t^* \in \operatorname{argmin}_{1\leq t\leq T}\|P(\mathbf{u}^t,\nabla\psi(\mathbf{u}^t))\|^2$. Pick iterates θ^{t^*} and θ^{t^*+1} of Algorithm 1. The corresponding iterates in the $\mathcal{U}$-space are $\mathbf{u}^{t^*} = \boldsymbol{\ell}(\theta^{t^*})$ and $\mathbf{u}^{t^*+1} = \boldsymbol{\ell}(\theta^{t^*+1})$.

Further, let $\tilde{\mathbf{u}}^{t^*+1} = \mathbf{u}^{t^*} - \eta\nabla\psi(\mathbf{u}^{t^*})$ be the un-projected next iterate, and $\hat{\mathbf{u}}^{t^*+1} = \mathbf{u}^{t^*} - \eta\cdot P(\mathbf{u}^{t^*},\nabla\psi(\mathbf{u}^{t^*}))$ be the one obtained after an exact projection, both using exact gradient $\nabla\psi(\mathbf{u}^{t^*})$.

We start with the assumption that (as promised by Lemma 6):

$$\mathbf{E}[\|P(\mathbf{u}^{t^*},\nabla\psi(\mathbf{u}^{t^*}))\|^2] = \frac{1}{\eta^2}\mathbf{E}[\|\mathbf{u}^{t^*} - \eta\cdot P(\mathbf{u}^{t^*},\nabla\psi(\mathbf{u}^{t^*})) - \mathbf{u}^{t^*}\|^2] = \frac{1}{\eta^2}\mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \mathbf{u}^{t^*}\|^2] \leq \epsilon^2$$

or equivalently,

$$\mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \mathbf{u}^{t^*}\|^2] \leq \eta^2\epsilon^2 \tag{10}$$

and would like to bound the gradient norm of $\psi(\boldsymbol{\ell}(\cdot))$ at θ^{t^*}.

We start by translating (10) to a guarantee in the θ-space. We know that

$$\hat{\mathbf{u}}^{t^*+1} \in \arg\min_{\mathbf{u}\in\mathcal{U}}\|\mathbf{u} - \tilde{\mathbf{u}}^{t^*+1}\|^2. \tag{11}$$

Put together (10) and (11), and take expectation over randomness in $\mathbf{u}^{t^*}$,

$$\begin{aligned}\mathbf{E}[\|\mathbf{u}^{t^*} - \tilde{\mathbf{u}}^{t^*+1}\|^2] &\leq \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \mathbf{u}^{t^*}\|^2] + 2\mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|\|\hat{\mathbf{u}}^{t^*+1} - \mathbf{u}^{t^*}\|]\\ &\leq \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \eta^2\epsilon^2 + 2\eta\epsilon\sqrt{\mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2]}\\ &\leq \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \eta^2\epsilon^2 + 2\eta\epsilon\sqrt{\mathbf{E}[\|\mathbf{u}^{t^*} - \tilde{\mathbf{u}}^{t^*+1}\|^2]}\\ &= \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \eta^2\epsilon^2 + 2\eta^2\epsilon\sqrt{\mathbf{E}[\|\nabla\psi(\mathbf{u}^{t^*})\|^2]},\end{aligned}$$

where we used Cauchy-Schwarz for the second step. Using the fact that ψ is L-Lipschitz:

$$\mathbf{E}[\|\mathbf{u}^{t^*} - \tilde{\mathbf{u}}^{t^*+1}\|^2] \leq \mathbf{E}[\|\hat{\mathbf{u}}^{t^*+1} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \epsilon', \tag{12}$$

where $\epsilon' = \eta^2(\epsilon^2 + 2L\epsilon)$.

We also know that $\hat{\mathbf{u}}^{t^*+1}$ can be equivalently obtained by performing an optimization in the θ-space as follows:

$$\hat{\theta}^{t^*+1} \in \arg\min_{\theta\in\mathbb{R}^d}\|\max\{\boldsymbol{\ell}(\theta),\tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2$$

and setting $\hat{\mathbf{u}}^{t^*+1} = \max\{\boldsymbol{\ell}(\hat{\theta}^{t^*+1}), \tilde{\mathbf{u}}^{t^*+1}\}$. So (12) translates to the following guarantee in the θ-space:

$$\mathbf{E}[\|\boldsymbol{\ell}(\theta^{t^*}) - \tilde{\mathbf{u}}^{t^*+1}\|^2] \leq \mathbf{E}[\min_{\theta\in\mathbb{R}^d} \|\max\{\boldsymbol{\ell}(\theta), \tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \epsilon', \tag{13}$$

where we have used $\mathbf{u}^{t^*} = \boldsymbol{\ell}(\theta^{t^*})$. Now since

$$\|\max\{\boldsymbol{\ell}(\theta^{t^*}), \tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2 = \|(\boldsymbol{\ell}(\theta^{t^*}) - \tilde{\mathbf{u}}^{t^*+1})_+\|^2 \leq \|\boldsymbol{\ell}(\theta^{t^*}) - \tilde{\mathbf{u}}^{t^*+1}\|^2,$$

together with (13) we have

$$\mathbf{E}[\|\max\{\boldsymbol{\ell}(\theta^{t^*}), \tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2] \leq \mathbf{E}[\min_{\theta\in\mathbb{R}^d} \|\max\{\boldsymbol{\ell}(\theta), \tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2] + \epsilon'. \tag{14}$$

Having translated our initial assumption on the gradient mapping to θ-space, we can now provide a guarantee on the gradient of $\psi(\boldsymbol{\ell}(\cdot))$. Let $Q(\theta) := \|\max\{\boldsymbol{\ell}(\theta), \tilde{\mathbf{u}}^{t^*+1}\} - \tilde{\mathbf{u}}^{t^*+1}\|^2 = \|(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^{t^*+1})_+\|^2$.

Taking as given that Q is smooth in θ with smoothness parameter ω for now, by standard properties of smooth functions, we have for any θ':

$$\|\nabla Q(\theta')\|^2 \leq 2\omega \cdot (Q(\theta') - \min_{\theta\in\mathbb{R}^d} Q(\theta)).$$

Using the above property and (14), taking expectation on both sides, we have:

$$\mathbf{E}[\|\nabla Q(\theta^{t^*})\|^2] \leq 2\omega\epsilon',$$

or equivalently,

$$\mathbf{E}\Big[\Big\|2\sum_{k=1}^{K}(\ell_k(\theta^{t^*}) - \tilde{u}_k^{t^*+1})_+\nabla_\theta\ell_k(\theta^{t^*})\Big\|^2\Big] \leq 2\omega\epsilon',$$

therefore

$$4\eta^2\mathbf{E}\Big[\Big\|\sum_{k=1}^{K}(\nabla\psi_k(\boldsymbol{\ell}^{t^*}))_+\nabla_\theta\ell_k(\theta^{t^*})\Big\|^2\Big] \leq 2\omega\epsilon',$$

where we use the short-hand $\boldsymbol{\ell}^{t^*} = \boldsymbol{\ell}(\theta^{t^*})$. By monotonicity of ψ, the gradient of ψ is always non-negative, and the above becomes:

$$4\eta^2\mathbf{E}\Big[\Big\|\sum_{k=1}^{K}\nabla\psi_k(\boldsymbol{\ell}^{t^*})\nabla_\theta\ell_k(\theta^{t^*})\Big\|^2\Big] \leq 2\omega\epsilon',$$

and we have:

$$\mathbf{E}[\|\nabla_\theta\psi(\boldsymbol{\ell}(\theta^{t^*}))\|^2] \leq \omega\epsilon'/2\eta^2 = \omega(\epsilon^2 + 2L\epsilon)/2,$$

as desired. It remains to justify the smoothness of $Q(\theta)$. For any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\begin{aligned}
&\|\nabla Q(\theta_1) - \nabla Q(\theta_2)\| \\
&= \Big\|2\sum_{k=1}^{K}(\ell_k(\theta_1) - \tilde{u}_k^{t^*+1})_+ \cdot \nabla_\theta\ell_k(\theta_1) - 2\sum_{k=1}^{K}(\ell_k(\theta_2) - \tilde{u}_k^{t^*+1})_+ \cdot \nabla_\theta\ell_k(\theta_2)\Big\| \\
&\leq 2\sum_{k=1}^{K}\Big\|(\ell_k(\theta_1) - \tilde{u}_k^{t^*+1})_+ \cdot (\nabla_\theta\ell_k(\theta_1) - \nabla_\theta\ell_k(\theta_2))\Big\| + \Big\|\big[(\ell_k(\theta_1) - \tilde{u}_k^{t^*+1})_+ - (\ell_k(\theta_2) - \tilde{u}_k^{t^*+1})_+\big] \cdot \nabla_\theta\ell_k(\theta_2)\Big\| \\
&\leq 2\sum_{k=1}^{K}|\ell_k(\theta_1) - \tilde{u}_k^{t^*+1}| \cdot \gamma\|\theta_1 - \theta_2\| + |\ell_k(\theta_1) - \ell_k(\theta_2)| \cdot \|\nabla_\theta\ell_k(\theta_2)\| \\
&= 2\sum_{k=1}^{K}|\ell_k(\theta_1) - \ell_k(\theta^{t^*}) + \eta\nabla\psi_k(u^{t^*})| \cdot \gamma\|\theta_1 - \theta_2\| + \Phi^2\|\theta_1 - \theta_2\| \\
&\leq 2K\big[(G + \eta L) \cdot \gamma + \Phi^2\big] \cdot \|\theta_1 - \theta_2\| = 2K\Big[(G + \frac{L}{\beta^2}) \cdot \gamma + \Phi^2\Big] \cdot \|\theta_1 - \theta_2\|
\end{aligned}$$

where we used γ-smoothness and Φ-lipschitz property of ℓ_k and $\|\boldsymbol{\ell}(\theta)\| \leq G$, together with $(a)_+ - (b)_+ \leq |a-b|$, therefore $\omega = 2K\big[(G + \frac{L}{\beta^2}) \cdot \gamma + \Phi^2\big]$. □

A.4. Proof of Lemma 3

Recall from Algorithm 2 that the finite difference estimate of the gradient of ψ at θ' is given by:

$$\hat{\mathbf{g}} = \frac{1}{m}\sum_{j=1}^{m} \frac{M(\mathbf{f}_{\theta'} + \Delta^j, \mathbf{y}) - M(\mathbf{f}_{\theta'}, \mathbf{y})}{\sigma} Z^j.$$

Lemma 3 (Restated). *Let M be as defined in Theorem 2 and $|\epsilon(\theta)| \leq \bar{\epsilon}, \forall \theta$. Let $\hat{\mathbf{g}}$ be returned by Algorithm 2 for a given θ', m perturbations and $\sigma = \frac{\sqrt{\bar{\epsilon}}}{\sqrt{K}\beta^2}$.*

$$\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2\right] \leq \mathcal{O}\left(\frac{L^2 K}{m} + \bar{\epsilon} K^2 \beta^2\right),$$

where the expectation is over the random perturbations.

We will find it useful to re-state results from Nesterov and Spokoiny (2017), extended to our setting.

Lemma 8. *Suppose ψ is L-Lipschitz and β-smooth. Define $\psi_\sigma(\mathbf{u}) := \mathbf{E}_{Z\sim\mathcal{N}(\mathbf{0},\mathbf{I}_K)}[\psi(\mathbf{u} + \sigma Z)]$. Let $\hat{\mathbf{g}}_1 = \frac{1}{m}\sum_{j=1}^{m} \frac{\psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta^j, \mathbf{y})) - \psi(\boldsymbol{\ell}(\mathbf{f}_\theta, \mathbf{y}))}{\sigma} Z^j$, where Δ^j is as defined in Algorithm 2. Then:*

1. *$\hat{\mathbf{g}}_1$ is an unbiased estimate of the gradient of ψ_σ at $\boldsymbol{\ell}(\theta)$, i.e., $\mathbf{E}[\hat{\mathbf{g}}_1] = \nabla\psi_\sigma(\boldsymbol{\ell}(\theta))$.*
2. $\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \mathbf{E}[\hat{\mathbf{g}}_1]\|^2\right] \leq \frac{\sigma^2\beta^2}{m}(K+6)^3 + \frac{4L^2}{m}(K+4)$.
3. $\|\nabla\psi_\sigma(\boldsymbol{\ell}(\theta)) - \nabla\psi(\boldsymbol{\ell}(\theta))\| \leq \frac{\sigma\beta^2}{2}(K+3)^{3/2}$.

Proof. See Eq. (21) in Nesterov et al. (2017) for part 1. Theorem 4 of Nesterov et al. together with the fact that $\mathrm{Var}(X) \leq \mathbf{E}[X^2]$ implies part 2. See Lemma 3 of Nesterov et al. for part 3. □

Proof of Lemma 3. We can write out the gradient estimate as:

$$\begin{aligned}
\hat{\mathbf{g}} &= \frac{1}{m}\sum_{j=1}^{m} \frac{M(\mathbf{f}_\theta + \Delta^j, \mathbf{y}) - M(\mathbf{f}_\theta, \mathbf{y})}{\sigma} Z^j \\
&= \frac{1}{m}\sum_{j=1}^{m} \frac{\psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta^j, \mathbf{y})) - \psi(\boldsymbol{\ell}(\mathbf{f}_\theta, \mathbf{y}))}{\sigma} Z^j + \frac{1}{m}\sum_{j=1}^{m} \frac{\epsilon(\mathbf{f}_\theta + \Delta^j, \mathbf{y}) - \epsilon(\mathbf{f}_\theta, \mathbf{y})}{\sigma} Z^j \\
&= \frac{1}{m}\sum_{j=1}^{m} \frac{\psi(\boldsymbol{\ell}(\theta) + \sigma Z^j) - \psi(\boldsymbol{\ell}(\theta))}{\sigma} Z^j + \frac{1}{m}\sum_{j=1}^{m} \frac{\epsilon(\mathbf{f}_\theta + \Delta^j, \mathbf{y}) - \epsilon(\mathbf{f}_\theta, \mathbf{y})}{\sigma} Z^j \\
&:= \hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2,
\end{aligned}$$

where $\epsilon(\mathbf{f}_\theta, \mathbf{y})$ is the unknown slack function in Section 3.1, re-written in terms of the scores $\mathbf{f}_\theta$ and labels $\mathbf{y}$.

Let ψ_σ be defined as in Lemma 8. Then the gradient estimate error can be expanded as:

$$\begin{aligned}
\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta))\|^2\right] &\leq 2\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi_\sigma(\boldsymbol{\ell}(\theta))\|^2\right] + 2\|\nabla\psi_\sigma(\boldsymbol{\ell}(\theta)) - \nabla\psi(\boldsymbol{\ell}(\theta))\|^2 \\
&\leq 4\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \nabla\psi_\sigma(\boldsymbol{\ell}(\theta))\|^2\right] + 4\mathbf{E}\left[\|\hat{\mathbf{g}}_2\|^2\right] + 2\|\nabla\psi_\sigma(\boldsymbol{\ell}(\theta)) - \nabla\psi(\boldsymbol{\ell}(\theta))\|^2 \\
&\leq 4\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \nabla\psi_\sigma(\boldsymbol{\ell}(\theta))\|^2\right] + \frac{16\bar{\epsilon}^2}{\sigma^2 m}\sum_{j=1}^{m}\mathbf{E}\left[\|Z^j\|^2\right] + 2\|\nabla\psi_\sigma(\boldsymbol{\ell}(\theta)) - \nabla\psi(\boldsymbol{\ell}(\theta))\|^2 \\
&\leq \frac{4\sigma^2}{m}\beta^2(K+6)^3 + \frac{16}{m}L^2(K+4) + \frac{16\bar{\epsilon}^2 K}{\sigma^2} + \frac{\sigma^2}{2}\beta^4(K+3)^3,
\end{aligned}$$

where we used the fact that (1) $\hat{\mathbf{g}}_1$ is an unbiased estimate of $\nabla\psi_\sigma(\boldsymbol{\ell}(\theta))$ (see part 1 of Lemma 8); (2) the assumption that $|\epsilon(\theta)| \leq \bar{\epsilon}$; (3) $\|a_1 + \cdots + a_m\|^2 \leq m(\|a_1\|^2 + \cdots + \|a_m\|^2)$, and the last step follows from Parts 2–3 of Lemma 8.

Setting $\sigma = \frac{\sqrt{\bar{\epsilon}}}{\sqrt{K}\beta^2}$ completes the proof. □

A.5. Proofs and Discussion for Linear Interpolation Gradient Estimates

Lemma 4 (Restated). *Let M be defined as in Theorem 2 and $|\epsilon(\theta)| \leq \bar{\epsilon}, \forall \theta$. Assume each ℓ_k is Φ-Lipschitz in θ w.r.t. the L_∞-norm, and $\|\boldsymbol{\ell}(\theta)\| \leq G \ \forall \theta$. Suppose for a given θ', σ and perturbation count m, the expected covariance matrix for the left-hand-side of the linear system $\mathbf{H}$ is well-conditioned with the smallest singlular value $\lambda_{\min}(\sum_{i=1}^m \mathbf{E}[\mathbf{H}_i \mathbf{H}_i^\top]) \geq \mu_{\min} = \mathcal{O}(m\sigma^2\Phi^2)$. Then for any $\delta > 0$, setting $\sigma = \frac{G^{1/3}\bar{\epsilon}^{1/3}}{\Phi K^{3/2}\log(d)^{2/3}\beta^{1/3}}$ and $m = \frac{G^4 K^9 \log(d)^4 \beta^2 \log(K/\delta)}{\bar{\epsilon}^2}$, Algorithm 3 returns w.p. $\geq 1-\delta$ (over draws of random perturbations) a gradient estimate $\hat{\mathbf{g}}$ that satisfies:*

$$\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 \leq \tilde{\mathcal{O}}\left(G^{1/3}\bar{\epsilon}^{1/3}K^3\beta^{2/3}\right).$$

We first discuss the assumptions in Lemma 4 in Section A.5.1. We then provide the proof for the high probability statement in the lemma in Section A.5.2. We then show how this can be translated to a bound on the expected gradient error via truncation in Section A.5.3.

A.5.1. Assumptions in Lemma 4

We discuss example settings where the assumptions in the lemma hold.

Correlation Assumption on H. One of the key assumptions we make is that the matrix $\mathbf{H}$ is well-conditioned. Recall that $\mathbf{H}$ is a $m \times K$ matrix, where each row corresponds to a perturbation of the surrogates, and contains differences in the K surrogates $\ell_1, \ldots, \ell_K$ at two independent perturbations to the model parameters θ. We assume that the smallest singular value of $\mathbf{H}$'s covariance matrix $\sum_{i=1}^m \mathbf{E}[\mathbf{H}_i\mathbf{H}_i^\top]$ scales as $m\sigma^2\Phi^2$. This assumption essentially states that the perturbations on the K surrogates are weakly correlated. The scaling factors σ and Φ come from the fact that Gaussian perturbations on the model parameters θ have standard deviation σ and the surrogates ℓ_k are Φ-Lipschitz.

As an example scenario where this assumption holds, consider a ML fairness task where the instances belong to K non-overlapping protected groups. Further, assume that the group membership attribute is included in the feature vector, i.e., the d-dimensional feature vector $\mathbf{x} = [g_1, \ldots, g_K, \tilde{x}_1, \ldots, \tilde{x}_{d-K}]$, where g_k is a Boolean indicating if the instance belongs to group k, and $\tilde{x}_1, \ldots, \tilde{x}_{d-K}$ are group-independent features. A natural choice of surrogates for this application would be average losses computed on the K individual groups. For example, with a linear model θ, we could choose ℓ_k to be the average squared loss conditioned on examples from group k, i.e., $\ell_k(\theta) = \mathbf{E}_{(x,y)|x_k=1}[(\theta^\top x - y)^2]$.

Note that the first K coordinates of the model vector θ correspond to weights on the K Boolean group attributes. So adding noise $Z_k \in \mathbb{R}$ to the k-th coordinate of θ only affects scores on examples from the k-th group (i.e., examples for which $x_k = 1$), and hence only perturbs surrogate ℓ_k. Specifically, adding $Z_k \in \mathbb{R}$ to the k-th coordinate of θ would perturb $\ell_k(\theta)$ to $\ell_k(\theta) + C_k Z_k + Z_k^2$, where $C_k = 2\mathbf{E}_{(x,y)|x_k=1}[\theta^\top x - y]$, and leave the other surrogates $\ell_j, j \neq k$ unchanged.

Now suppose we add independent σ-Gaussian noise to only the first K coordinates of θ. The expected covariance matrix as defined in the lemma statement then takes the form:

$$\sum_{i=1}^m \mathbf{E}[\mathbf{H}_i\mathbf{H}_i^\top] = \begin{bmatrix} \mathcal{O}(m(C_1^2\sigma^2 + \sigma^4)) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \mathcal{O}(m(C_K^2\sigma^2 + \sigma^4)) \end{bmatrix} = \begin{bmatrix} \Omega(m\sigma^2) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \Omega(m\sigma^2) \end{bmatrix},$$

where recall that the k-th column of $\mathbf{H}$ contains the differences of $\ell_k(\theta)$ at two different σ-Gaussian perturbations on the first K coordinates of θ, and C_k's are constants that are independent of the random perturbations.

In the more general case, where we perturb all coordinates of θ, the assumption on $\mathbf{H}$ would still hold if there exists a subset of coordinates for each surrogate ℓ_k that when perturbed produce larger changes to ℓ_k than to the other surrogates.

Lipschitz Assumption on $\boldsymbol{\ell}(\theta)$ Another key assumption we make is that the surrogates ℓ_k are Φ-Lipschitz w.r.t. the L_∞-norm. This allows us to produce perturbations in the K surrogates by perturbing the model parameters θ, and do so without a strong dependence on the dimension of θ in the error bound. Note that the choice of the infinity norm results in a mild logarithmic dependence on the dimension d in the bound. When the surrogates the are not L_∞-Lipschitz, but are instead Lipschitz w.r.t. the L_2-norm, we prescribe perturbing only a small number of $d' \ll d$ coordinates of θ that are most closely related to the surrogate (such as e.g. the group attribute coordinates in the fairness example above), and this would result in a bound that has a polynomial dependence on d'.

A.5.2. PROOF OF LEMMA 4

We will make use of the fact that because we perturb the model parameters θ with Gaussian random noise, the resulting perturbations on the surrogates $\boldsymbol{\ell}$ follow a sub-Gaussian distribution. We first state a few well-known facts about sub-Gaussian random vectors.

Lemma 9 (Properties of sub-Gaussian distribution)**.**

(i) Let $(Z_1, \cdots, Z_d)$ be a vector of i.i.d standard gaussian variables and $f : \mathbb{R}^N \to \mathbb{R}$ be Φ-Lipschitz w.r.t. L_2-norm. Then the random variable $f(\sigma Z) - \mathbf{E}[f(\sigma Z)]$ is sub-Gaussian with parameter at most $\sigma\Phi$.

(ii) Let $Z_1, \cdots, Z_K$ be K (not necessarily independent) sub-Gaussian random variables with parameters at most σ. Then the random vector $(Z_1, \cdots, Z_K)$ is a sub-Gaussian random vector with parameter σK.

(iii) For a sub-Gaussian random vector $Z \in \mathbb{R}^K$ with parameter at most σ, we have for any $p \in \mathbb{N}$:

$$(\mathbf{E}[\|Z - \mathbf{E}[Z]\|_2^p])^{1/p} \leq 2\sqrt{2}\sigma\sqrt{K}\sqrt{p}.$$

Proof. For a proof of (1), see e.g. Wainwright (2019), Chapter 2. For a proof of (3), see Jin et al. (2019). We now prove (2).

For a random vector $(Z_1, \ldots, Z_K)$ where the coordinates Z_k's are σ-sub-Gaussian and not necessarily independent, we have that for any $v \in \mathbb{S}^{K-1}$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbf{E}[\exp(\lambda v^\top (Z - \mathbf{E}[Z]))] &= \mathbf{E}\Big[\prod_{k=1}^K \exp\Big(\lambda v_k (Z_k - \mathbf{E}[Z_k])\Big)\Big] \\
&\leq \prod_{k=1}^K \mathbf{E}\Big[\Big(\exp(\lambda v_k (Z_k - \mathbf{E}[Z_k]))\Big)^K\Big]^{1/K} \\
&\leq \prod_{k=1}^K \exp\left(\frac{1}{2}\lambda^2 K^2 \sigma^2\right)^{1/K} = \prod_{k=1}^K \exp\left(\frac{1}{2}\lambda^2\sigma^2 K\right) \leq \exp\left(\frac{1}{2}\lambda^2\sigma^2 K^2\right),
\end{aligned}$$

where we have used Hölder's inequality for the second step. □

We can write the optimization problem in Algorithm 3 as solving the following linear system

$$\begin{bmatrix} h'_{11} - h''_{11} & \cdots & h'_{1K} - h''_{1K} \\ \vdots & \cdots & \vdots \\ h'_{m1} - h''_{m1} & \cdots & h'_{mK} - h''_{mK} \end{bmatrix} \cdot \hat{\mathbf{g}} = \begin{bmatrix} \psi(\boldsymbol{\ell}(\theta') + \mathbf{h}'_1) - \psi(\boldsymbol{\ell}(\theta') + \mathbf{h}''_1) + \epsilon_{11} - \epsilon_{12} \\ \vdots \\ \psi(\boldsymbol{\ell}(\theta') + \mathbf{h}'_m) - \psi(\boldsymbol{\ell}(\theta') + \mathbf{h}''_m) + \epsilon_{m1} - \epsilon_{m1} \end{bmatrix},$$

and use the resulting $\hat{\mathbf{g}} \in \mathbb{R}^K$ as the gradient estimate, where we denote $\mathbf{h}'_j := \boldsymbol{\ell}(\theta' + \sigma Z_1^j) - \boldsymbol{\ell}(\theta') \in \mathbb{R}^K$ and $\mathbf{h}''_j := \boldsymbol{\ell}(\theta' + \sigma Z_2^j) - \boldsymbol{\ell}(\theta') \in \mathbb{R}^K$ for $j \in [m]$, and $\epsilon_{j1} = \epsilon(\theta' + \sigma Z_1^j)$ and $\epsilon_{j2} = \epsilon(\theta' + \sigma Z_2^j)$. We further denote

$$\begin{gathered}
\mathbf{L} = [\boldsymbol{\ell}(\theta'); \ldots; \boldsymbol{\ell}(\theta')] \in \mathbb{R}^{m \times K} \\
\mathbf{H}' = [\mathbf{h}'_1; \cdots; \mathbf{h}'_m] \in \mathbb{R}^{m \times K}, \quad \mathbf{H}'' = [\mathbf{h}''_1; \cdots \mathbf{h}''_m] \in \mathbb{R}^{m \times K} \\
\boldsymbol{\epsilon}_1 = [\epsilon_{11}; \ldots; \epsilon_{m1}] \in \mathbb{R}^m, \quad \boldsymbol{\epsilon}_2 = [\epsilon_{12}; \ldots; \epsilon_{m2}] \in \mathbb{R}^m,
\end{gathered}$$

and equivalently re-write the above linear system as:

$$(\mathbf{H}' - \mathbf{H}'') \cdot \hat{\mathbf{g}} = \psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') + \boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2, \tag{15}$$

where the matrix $\mathbf{H}$ that we defined in the lemma statement is the same as $\mathbf{H}' - \mathbf{H}''$.

Below we state a lemma involving implications of our assumptions on the left-hand-side perturbation matrices $\mathbf{H}'$ and $\mathbf{H}''$.

Lemma 10 (Properties of $\mathbf{H}'$ and $\mathbf{H}''$)**.** *Suppose each $\ell_k(\theta)$ is Φ-Lipschitz w.r.t. the L_∞-norm and $\|\boldsymbol{\ell}(\theta)\| \leq G, \forall\theta$. Then each $\mathbf{h}'_i$ and each $\mathbf{h}''_i$ is a sub-Gaussian vector with parameter at most $\sigma\Phi K$. The differences $\mathbf{h}'_i - \mathbf{h}''_i$ are also sub-Gaussian random vectors with parameter at most $2\sigma\Phi K$, and have mean zero. Moreover, $\|\mathbf{h}'_i\| \leq 2G$, $\|\mathbf{h}''_i\| \leq 2G$, $\|\mathbf{h}'_i - \mathbf{h}''_i\| \leq 2G$, $\forall i$.*

The proof follows directly from Lemma 9(i)–(ii) and the fact that a function ℓ_k that is Φ-Lipschitz w.r.t. the L_∞-norm is also Φ-Lipschitz w.r.t. the L_2-norm. We also have from the smoothness of ψ that

$$\left|\psi(\boldsymbol{\ell}(\theta') + \mathbf{h}'_i) - [\psi(\boldsymbol{\ell}(\theta')) + \nabla\psi(\boldsymbol{\ell}(\theta'))^\top \mathbf{h}'_i]\right| \leq \frac{\beta}{2}\|\mathbf{h}'_i\|_2^2. \tag{16}$$

With this in hand, we are ready to bound the error in the gradient estimate $\hat{\mathbf{g}}$ compared to $\nabla\psi(\boldsymbol{\ell}(\theta))$.

Proof of Lemma 4. The least squares estimate for the linear system in (15) is given by:

$$\begin{aligned}
\hat{\mathbf{g}} &= \Big((\mathbf{H}' - \mathbf{H}'')^\top(\mathbf{H}' - \mathbf{H}'')\Big)^{-1}(\mathbf{H}' - \mathbf{H}'')^\top[\psi(\mathbf{L} + \mathbf{H}') + \boldsymbol{\epsilon_1} - \psi(\mathbf{L} + \mathbf{H}'') - \boldsymbol{\epsilon_2}] \\
&= \Big((\mathbf{H}' - \mathbf{H}'')^\top(\mathbf{H}' - \mathbf{H}'')\Big)^{-1}(\mathbf{H}' - \mathbf{H}'')^\top\Big[(\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') \\
&\qquad\qquad - (\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\Big] \\
&= \nabla\psi(\boldsymbol{\ell}(\theta')) + \Big((\mathbf{H}' - \mathbf{H}'')^\top(\mathbf{H}' - \mathbf{H}'')\Big)^{-1}(\mathbf{H}' - \mathbf{H}'')^\top\Big[\psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') \\
&\qquad\qquad - (\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\Big].
\end{aligned}$$

The error in the least squares based gradient estimate is then:

$$\begin{aligned}
&\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\| \\
&\leq \underbrace{\left\|\Big((\mathbf{H}' - \mathbf{H}'')^\top(\mathbf{H}' - \mathbf{H}'')\Big)^{-1}\right\|_{\text{op}}}_{\text{term}_1}\underbrace{\left\|(\mathbf{H}' - \mathbf{H}'')^\top\Big[\psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') - (\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\Big]\right\|_2}_{\text{term}_2}.
\end{aligned} \tag{17}$$

Bounding the second term in (17). We first bound the second term in (17). We have:

$$\begin{aligned}
&\|\psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') - (\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\|_2 \\
&= \left\|\psi(\mathbf{L} + \mathbf{H}') - \mathbf{H}'\nabla\psi(\boldsymbol{\ell}) - \psi(\mathbf{L}) + \psi(\mathbf{L}) + \mathbf{H}''\nabla\psi(\boldsymbol{\ell}) - \psi(\mathbf{L} + \mathbf{H}'') + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\right\|_2 \\
&\leq \left\|\psi(\mathbf{L} + \mathbf{H}') - \mathbf{H}'\nabla\psi(\boldsymbol{\ell}) - \psi(\mathbf{L})\right\|_2 + \left\|\psi(\mathbf{L}) + \mathbf{H}''\nabla\psi(\boldsymbol{\ell}) - \psi(\mathbf{L} + \mathbf{H}'')\right\|_2 + \|\boldsymbol{\epsilon_1}\|_2 + \|\boldsymbol{\epsilon_2}\|_2 \\
&\leq \frac{\beta}{2}\|\mathbf{H}'\|_{\text{F}}^2 + \frac{\beta}{2}\|\mathbf{H}''\|_{\text{F}}^2 + 2\sqrt{m}\bar{\epsilon},
\end{aligned}$$

where we used (16) and the the assumption $|\epsilon(\theta)| \leq \bar{\epsilon}\ \forall\theta$. This in turn gives

$$\begin{aligned}
\text{term}_2 \ &= \ \left\|(\mathbf{H}' - \mathbf{H}'')^\top\Big[\psi(\mathbf{L} + \mathbf{H}') - \psi(\mathbf{L} + \mathbf{H}'') - (\mathbf{H}' - \mathbf{H}'')\nabla\psi(\boldsymbol{\ell}) + \boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}\Big]\right\|_2 \\
&\leq \sum_{j=1}^{m}\|\mathbf{h}'_j - \mathbf{h}''_j\| \cdot \frac{\beta}{2}(\|\mathbf{h}'_j\|_2^2 + \|\mathbf{h}''_j\|_2^2) + 2\sqrt{m}G\sqrt{m}\bar{\epsilon}
\end{aligned}$$

where each $\mathbf{h}'_j$ is of length K with (correlated) subgaussian coordinates. Therefore using Cauchy-Schwarz,

$$\mathbf{E}[\|\mathbf{h}'_j - \mathbf{h}''_j\| \cdot \|\mathbf{h}'_j\|_2^2] \leq \sqrt{\mathbf{E}[\|\mathbf{h}'_j - \mathbf{h}''_j\|_2^2]} \cdot \sqrt{\mathbf{E}[\|\mathbf{h}'_j\|_2^4]}.$$

Note that $\mathbf{E}[h'_{ij}] = \mathbf{E}[\ell_j(\theta' + \sigma Z^i)] - \ell_j(\theta') \leq \sigma\Phi\mathbf{E}[\|Z^i\|_\infty] \leq \mathcal{O}(\sigma\Phi\sqrt{\log(d)})$,where we've used that the max of d independent standard normal random variables scales as $\sqrt{\log(d)}$. Similarly, $\mathbf{E}[h''_{ij}] \leq \mathcal{O}(\sigma\Phi\sqrt{\log(d)})$. Together with these facts and Lemma 10 and Lemma 9(iii) we have

$$\sqrt{\mathbf{E}[\|\mathbf{h}'_j\|_2^4]} \leq \sqrt{8(4\sqrt{2}\sigma\Phi K^{3/2})^4 + \mathcal{O}(\sigma^2\Phi^2\log(d)K)^2} \leq \mathcal{O}(\sigma^2\Phi^2K^3(\log(d))^2)$$

where we used triangle inequality and $(a+b)^p \le 2^{p-1}(a^p+b^p)$. Similarly, we have:

$$\sqrt{\mathbf{E}[\|\mathbf{h}'_j - \mathbf{h}''_j\|_2^2]} \le 4\sigma\Phi K^{3/2}\,.$$

Now since $\|\mathbf{h}'_j\|_2 \le G$, we can apply Hoeffding's inequality to these bounded random variables to get

$$\mathbf{P}\left(\sum_{j=1}^{m} \|\mathbf{h}'_j - \mathbf{h}''_j\|_2 \cdot \|\mathbf{h}'_j\|_2^2 \;\ge\; \mathcal{O}(\sigma^3\Phi^3K^{9/2}(\log(d))^2 m) \;+\; mt\right) \le 2\exp\left(-\frac{2mt^2}{G^6}\right),$$

which further gives us:

$$\mathbf{P}\left(\text{term}_2 \;\ge\; \mathcal{O}(\sigma^3\Phi^3K^{9/2}(\log(d))^2 m\beta) \;+\; m\beta t \;+\; 2mG\bar{\epsilon}\right) \le 2\exp\left(-\frac{2mt^2}{G^6}\right), \tag{18}$$

Bounding the first term in (17). Now the first term in (17) is simply

$$\text{term}_1 \;=\; \left\|\Big((\mathbf{H}'-\mathbf{H}'')^\top(\mathbf{H}'-\mathbf{H}'')\Big)^{-1}\right\|_{\text{op}} = \lambda_{\min}^{-1}\Big((\mathbf{H}'-\mathbf{H}'')^\top(\mathbf{H}'-\mathbf{H}'')\Big)\,.$$

Let us denote $\hat{\mathbf{\Sigma}} := \sum_{i=1}^m (\mathbf{h}'_i - \mathbf{h}''_i)(\mathbf{h}'_i - \mathbf{h}''_i)^\top$ as the empirical covariance matrix. We now apply a matrix Chernoff inequality (see e.g. Tropp (2015)) to lower bound the smallest eigenvalue of $\hat{\mathbf{\Sigma}}$. We first note that the largest eigenvalue of this matrix is bounded above:

$$\lambda_{\max}(\hat{\mathbf{\Sigma}}) = \max_{\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}\left((\mathbf{h}'_i - \mathbf{h}''_i)^\top u\right)^2 \le 4G^2\,,$$

This together with the matrix Chernoff bound gives us for $\mu_{\min} \le \lambda_{\min}(\hat{\mathbf{\Sigma}})$, we have

$$\mathbf{P}\left(\lambda_{\min}(\hat{\mathbf{\Sigma}}) \le \frac{\mu_{\min}}{2}\right) \le K\cdot\exp\Big(-\frac{\mu_{\min}}{32G^2}\Big).$$

The assumption $\mu_{\min} = \mathcal{O}(m\sigma^2\Phi^2)$ then yields:

$$\mathbf{P}\left(\text{term}_1 \le \mathcal{O}(m\sigma^2\Phi^2)\right) \le K\cdot\exp\Big(-\frac{m\sigma^2\Phi^2}{G^2}\Big). \tag{19}$$

Combining the above bound (19) with the bound on the second term (18) (picking $t = \sigma^3\Phi^3$), we get the following tail bound:

$$\mathbf{P}\left(\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\| \ge \mathcal{O}\left(\sigma\Phi K^{9/2}\log(d)^2\beta + \frac{G\bar{\epsilon}}{\sigma^2\Phi^2}\right)\right) \le K\cdot\exp\Big(-\frac{m\sigma^2\Phi^2}{G^2}\Big) + 4\exp\Big(-\frac{2m\sigma^6\Phi^6}{G^6}\Big)\,.$$

Then for any $\delta > 0$, setting $\sigma = \dfrac{G^{1/3}\bar{\epsilon}^{1/3}}{\Phi K^{3/2}\log(d)^{2/3}\beta^{1/3}}$ and $m = \dfrac{G^4K^9\log(d)^4\beta^2\log(K/\delta)}{\bar{\epsilon}^2}$, Algorithm 3 returns w.p. $\ge 1-\delta$ (over draws of random perturbations) a gradient estimate $\hat{\mathbf{g}}$ that satisfies:

$$\|\hat{\mathbf{g}} \;-\; \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 \;\le\; \mathcal{O}\left(G^{1/3}\bar{\epsilon}^{1/3}K^3(\log(d))^{4/3}\beta^{2/3}\right),$$

which completes the proof. □

A.5.3. Translating to a Bound on the Expected Error

Lemma 4 provides a high probability bound on the gradient estimation error. This means that with a small probability the gradient estimation error may not be bounded. To translate this high probability bound into a bound on the expected gradient error, we first truncate the estimated gradients to be in a bounded range:

$$\text{trunc}(\hat{\mathbf{g}}) \;=\; \begin{cases} \hat{\mathbf{g}} & \text{if } \|\hat{\mathbf{g}}\| \le 2\sqrt{K}L \\ \mathbf{0} & \text{otherwise} \end{cases},$$

where L is the Lipschitz constant for ψ.

Corollary 1. *Under the assumptions in Lemma 4, for any $\delta \in (0,1)$, setting $\sigma = \frac{G^{1/3}\bar{\epsilon}^{1/3}}{\Phi K^{3/2}\log(d)^{2/3}\beta^{1/3}}$ and $m = \frac{G^4K^9\log(d)^4\beta^2\log(K/\delta)}{\bar{\epsilon}^2}$, Algorithm 3 returns a gradient estimate $\hat{\mathbf{g}}$ that satisfies:*

$$\mathbf{E}\left[\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2\right] \leq \tilde{\mathcal{O}}\left(G^{1/3}\bar{\epsilon}^{1/3}K^3\beta^{2/3}\right) + 10KL^2\delta.$$

Proof. Because both the truncated gradient estimates and the true gradients are bounded, the gradient error is trivially bounded by:

$$\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 \leq 2(\|\text{trunc}(\hat{\mathbf{g}})\|^2 + \|\nabla\psi(\boldsymbol{\ell}(\theta'))\|^2) \leq 2(4KL^2 + L^2) \leq 10KL^2. \tag{20}$$

In the case where $\|\hat{\mathbf{g}}\| \leq 2\sqrt{K}L$, the gradient error for the truncated $\hat{\mathbf{g}}$ is the same as that for $\hat{\mathbf{g}}$:

$$\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 = \|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2. \tag{21}$$

When $\|\hat{\mathbf{g}}\| > 2\sqrt{K}L$, the gradient error for the truncated estimates $\text{trunc}(\hat{\mathbf{g}})$ is upper bounded by:

$$\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 = \|\nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 \leq L^2,$$

whereas the the gradient error for the original estimates $\hat{\mathbf{g}}$ is lower bounded by:

$$\begin{aligned}\|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 &\geq \max_{k\in[K]}\left(\hat{g}_k - \nabla_k\psi(\boldsymbol{\ell}(\theta'))\right)^2 \geq \left(\max_{k\in[K]}|\hat{g}_k| - \max_{k\in[K]}|\nabla_k\psi(\boldsymbol{\ell}(\theta'))|\right)^2 \\ &\geq \left(\frac{1}{\sqrt{K}}(2\sqrt{K}L) - L\right)^2 = L^2\,.\end{aligned}$$

Therefore even in this case, the gradient error for $\text{trunc}(\hat{\mathbf{g}})$ is bounded by that for $\hat{\mathbf{g}}$:

$$\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2 \leq L^2 \leq \|\hat{\mathbf{g}} - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2. \tag{22}$$

Combining (21) and (22) with the trivial upper bound in (20) allows us to convert the high probability result in Lemma 4 to the following bound on the expected error. For any $\delta \in (0,1)$, setting $\sigma = \frac{G^{1/3}\bar{\epsilon}^{1/3}}{\Phi K^{3/2}\log(d)^{2/3}\beta^{1/3}}$ and $m = \frac{G^4K^9\log(d)^4\beta^2\log(K/\delta)}{\bar{\epsilon}^2}$, we have:

$$\mathbf{E}\left[\|\text{trunc}(\hat{\mathbf{g}}) - \nabla\psi(\boldsymbol{\ell}(\theta'))\|^2\right] \leq \tilde{\mathcal{O}}\left((1-\delta)G^{1/3}\bar{\epsilon}^{1/3}K^3\beta^{2/3}\right) + 10\delta KL^2,$$

as desired. □

B. Handling Non-smooth Metrics

For ψ that is only L-Lipschitz and non-smooth, we extend the finite difference gradient estimate in Section 5.1 with a two-step perturbation method, as detailed in Algorithm 4. This approach can be seen as computing a finite-difference gradient estimate for a smooth approximation to the original ψ, given by $\psi_{\sigma_1}(\mathbf{u}) := \mathbf{E}\left[\psi(\mathbf{u} + \sigma_1 Z_1)\right]$, where $Z_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$. Since ψ_{σ_1} is a convolution of ψ with a Gaussian density kernel, it is always smooth. For this setting, we build on recent work by Duchi et al. (2015), and show that the two-step perturbation approach provides a gradient estimate for ψ_{σ_1}.

Lemma 11 (Two-step finite difference gradient estimate)**.** *Let $M(\theta) = \psi(\boldsymbol{\ell}(\theta)) + \epsilon(\theta)$, for a ψ that is L-Lipschitz, and the worst-case slack $\max_{\theta\in\mathbb{R}^d}|\epsilon(\theta)|$ is the minimum among all such decompositions of M. Suppose $|\epsilon(\theta)| \leq \bar{\epsilon}$, $\forall\theta$. Let $\hat{\mathbf{g}}$ be returned by Algorithm 4 for a fixed $\sigma_1 > 0$ and $\sigma_2 = \sqrt{\frac{\sigma_1}{K^{3/2}L}}$. Then:*

$$\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{L^{7/4}K^{13/8}}{m\sigma_1^{1/4}} + \frac{LK^{5/2}\bar{\epsilon}^2}{\sigma_1}\right).$$

Drawing upon the result of Theorem 2, we can repeat the analysis on the smooth function $\psi_{\sigma_1}(\cdot)$ to get the following convergence guarantee for Algorithm 1.

Algorithm 4 Two-step Finite-difference Gradient Estimate

Input: $\theta \in \mathbb{R}^d, M, \ell_1, \cdots, \ell_k$, estimation accuracy ϵ
Draw $Z_1^1, \ldots, Z_1^m, Z_2^1, \ldots, Z_2^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$
Find $\Delta_1^j \in \mathbb{R}^n$ s.t. $\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y}) = \boldsymbol{\ell}(\mathbf{f}_\theta, \mathbf{y}) + \sigma_1 Z_1^j$ for $j = 1, \ldots, m$
Find $\Delta_2^j \in \mathbb{R}^n$ s.t. $\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y}) = \boldsymbol{\ell}(\mathbf{f}_\theta, \mathbf{y}) + \sigma_1 Z_1^j + \sigma_2 Z_2^j$ for $j = 1, \ldots, m$
$\hat{\mathbf{g}} = \frac{1}{m} \sum_{j=1}^{m} \frac{M(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y}) - M(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y})}{\sigma_2} Z_2^j$
Output: $\hat{\mathbf{g}}$

Corollary 2 (Convergence of Algorithm 1 for non-smooth ψ)**.** *Let $M(\theta) = \psi(\boldsymbol{\ell}(\theta)) + \epsilon(\theta)$, for a ψ that is monotonic, and L-Lipschitz, and the worst-case slack $\max_{\theta \in \mathbb{R}^d} |\epsilon(\theta)|$ is the minimum among all such decompositions of M.*

Suppose each ℓ_k is γ-smooth and Φ-Lipschitz in θ with $\|\boldsymbol{\ell}(\theta)\| \leq G, \forall \theta$. Suppose the gradient $\hat{\mathbf{g}}^t$ are estimated with Algorithm 1 for a choice $\sigma_1 > 0$, number of perturbation m, and $\sigma_2 = \sqrt{\frac{\sigma_1}{K^{3/2}L}}$. Suppose the projection step satisfies $\|(\boldsymbol{\ell}(\theta^{t+1}) - \tilde{\mathbf{u}}^t)_+\|^2 \leq \min_{\theta \in \mathbb{R}^d} \|(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^t)_+\|^2 + \mathcal{O}(\frac{\sigma_1^2}{TKL^2}), \forall t \in [T]$. Set stepsize $\eta = \frac{\sigma_1^2}{KL^2}$.

Then Algorithm 1 converges to an approximate stationary point of the smooth approximation $\psi_{\sigma_1}(\boldsymbol{\ell}(\cdot))$:

$$\min_{1 \leq t \leq T} \mathbf{E}\left[\|\nabla \psi_{\sigma_1}(\boldsymbol{\ell}(\theta^t))\|^2\right] \leq C\left(\frac{\sqrt{K}L}{\sigma_1\sqrt{T}} + \sqrt{\kappa} + \sqrt{L}\kappa^{1/4}\right),$$

where the expectation is over the randomness in the gradient estimates, and $C = \mathcal{O}\big(KL\big(\gamma\big(G + \frac{\sigma_1^2}{KL}\big) + \Phi^2\big)\big)$ and $\kappa = \tilde{\mathcal{O}}\Big(\frac{L^{7/4}K^{13/8}}{m\sigma_1^{1/4}} + \frac{LK^{5/2}\bar{\epsilon}^2}{\sigma_1}\Big)$.

The above result guarantees convergence to the stationary point of the smoothed metric $\psi_{\sigma_1}(\boldsymbol{\ell}(\cdot))$ and not the original metric $\psi(\boldsymbol{\ell}(\cdot))$. However, as long as the surrogate functions $\boldsymbol{\ell}$ are continuously differentiable, by taking $\sigma_1 \to 0$ and allowing T to increase as σ_1 decreases, the algorithm can be made to converge to a stationary point of the original metric $\psi(\boldsymbol{\ell}(\cdot))$, in the sense of Clark-subdifferential (see e.g. Garmanjani and Vicente (2013)).

B.1. Proof of Lemma 11

We will find it useful to re-state results from Duchi et al. (2015) and Nesterov and Spokoiny (2017), extended to our setting.

Lemma 12. *Suppose ψ is L-Lipschitz. Define $\psi_{\sigma_1}(\mathbf{u}) := \mathbf{E}_{Z_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)}[\psi(\mathbf{u} + \sigma_1 Z_1)]$ and $\psi_{\sigma_1,\sigma_2}(\mathbf{u}) := \mathbf{E}_{Z_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)}[\psi_{\sigma_1}(\mathbf{u} + \sigma_2 Z_2)]$. Let $\hat{\mathbf{g}}_1 = \frac{1}{m}\sum_{j=1}^m \frac{\psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y})) - \psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y}))}{\sigma_2} Z_2^j$, where Δ_1^j, Δ_2^j are as defined in Algorithm 4. Then:*

1. *$\hat{\mathbf{g}}_1$ is an unbiased estimate of the gradient of ψ_{σ_1,σ_2} at $\boldsymbol{\ell}(\theta)$, i.e., $\mathbf{E}[\hat{\mathbf{g}}_1] = \nabla \psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta))$.*
2. *$\psi_{\sigma_1}(\cdot)$ is smooth with smoothness parameter $\frac{\sqrt{K}L}{\sigma_1}$ and Lipschitz with constant L.*
3. *$\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \mathbf{E}[\hat{\mathbf{g}}_1]\|^2\right] \leq \frac{CL^2K}{m}\left(\sqrt{\frac{\sigma_2}{\sigma_1}}K + \log K + 1\right)$ for some constant C.*
4. *$\|\nabla \psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta)) - \nabla \psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\| \leq \frac{\sigma_2}{2}\frac{\sqrt{K}L}{\sigma_1}(K+3)^{\frac{3}{2}}$.*

Proof. Part 1 follows by trivially observing

$$\mathbf{E}_{Z_1,Z_2}[\hat{\mathbf{g}}_1] = \mathbf{E}_{Z_2}\Big[\frac{\psi_{\sigma_1}(\mathbf{u} + \sigma_2 Z_2) - \psi_{\sigma_1}(\mathbf{u})}{\sigma_2} Z_2\Big] = \nabla \psi_{\sigma_1,\sigma_2}(\mathbf{u})$$

where we invoked part 1 of Lemma 8. See Lemma 2 of Nesterov and Spokoiny (2017) for part 2. Part 2 together with Lemma 2 in Duchi et al. (2015) give the result in part 3. See Lemma 3 of Nesterov and Spokoiny (2017) for part 4. □

Now we are ready to bound the MSE in gradient estimate.

Proof of Lemma 11. We can write out the gradient estimate as:

$$\begin{aligned}
\hat{\mathbf{g}} &= \frac{1}{m}\sum_{j=1}^{m}\frac{M(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y}) - M(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y})}{\sigma_2} Z_2^j \\
&= \frac{1}{m}\sum_{j=1}^{m}\frac{\psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y})) - \psi(\boldsymbol{\ell}(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y}))}{\sigma_2} Z_2^j + \frac{1}{m}\sum_{j=1}^{m}\frac{\epsilon(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y}) - \epsilon(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y})}{\sigma_2} Z_2^j \\
&= \frac{1}{m}\sum_{j=1}^{m}\frac{\psi(\boldsymbol{\ell}(\theta) + \sigma_1 Z_1^j + \sigma_2 Z_2^j) - \psi(\boldsymbol{\ell}(\theta) + \sigma_1 Z_1^j)}{\sigma_2} Z_2^j + \frac{1}{m}\sum_{j=1}^{m}\frac{\epsilon(\mathbf{f}_\theta + \Delta_2^j, \mathbf{y}) - \epsilon(\mathbf{f}_\theta + \Delta_1^j, \mathbf{y})}{\sigma_2} Z_2^j \\
&:= \hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2,
\end{aligned}$$

where $\epsilon(\mathbf{f}_\theta, \mathbf{y})$ is the unknown slack function in Section 3.1, re-written in terms of the scores $\mathbf{f}_\theta$ and labels $\mathbf{y}$.

Let ψ_{σ_1} and ψ_{σ_1,σ_2} be defined as in Lemma 12. Then the gradient estimate error can be expanded as:

$$\begin{aligned}
\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\|^2\right] &\le 2\mathbf{E}\left[\|\hat{\mathbf{g}} - \nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta))\|^2\right] + 2\|\nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta)) - \nabla\psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\|^2 \\
&\le 4\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta))\|^2\right] + 4\mathbf{E}\left[\|\hat{\mathbf{g}}_2\|^2\right] + 2\|\nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta)) - \nabla\psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\|^2 \\
&\le 4\mathbf{E}\left[\|\hat{\mathbf{g}}_1 - \nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta))\|^2\right] + \frac{16\bar{\epsilon}^2}{m\sigma_2^2}\sum_{j=1}^{m}\mathbf{E}\left[\|Z_2^j\|^2\right] + 2\|\nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta)) - \nabla\psi_{\sigma_1}(\boldsymbol{\ell}(\theta))\|^2 \\
&\le \frac{CL^2K}{m}\left(\sqrt{\frac{\sigma_2}{\sigma_1}}K + \log K + 1\right) + \frac{16\bar{\epsilon}^2K}{\sigma_2^2} + \frac{\sigma_2^2}{2}\frac{KL^2}{\sigma_1^2}(K+3)^3,
\end{aligned}$$

where we used that (1) $\hat{\mathbf{g}}_1$ is an unbiased estimate of $\nabla\psi_{\sigma_1,\sigma_2}(\boldsymbol{\ell}(\theta))$ (see part 1 of Lemma 12); (2) boundness assumption $|\epsilon(\theta)| \le \bar{\epsilon}$; (3) $\|a_1 + \cdots + a_m\|^2 \le m(\|a_1\|^2 + \cdots + \|a_m\|^2)$, and the last step follows from Parts 3–4 of Lemma 12.

Setting $\sigma_2 = \sqrt{\frac{\sigma_1}{K^{3/2}L}}$ completes the proof. □

B.2. Proof of Corollary 2

Proof. We begin by observing that convolution operation preserves monotonicity, convexity, and range of the function. Let $g_{\sigma_1}(\cdot)$ denotes Gaussian density function with variance σ_1^2, since $\psi_{\sigma_1}(\mathbf{u})$ is a positively-weighted linear combination of shifted $\psi(\cdot)$, i.e.,

$$\psi_{\sigma_1}(\mathbf{u}) = \int_{\mathbb{R}^K}\psi(\mathbf{u}-\mathbf{z})\cdot g_{\sigma_1}(\mathbf{z})\,d\mathbf{z} = \int_{\mathbb{R}^K}\psi(\mathbf{z})\cdot g_{\sigma_1}(\mathbf{u}-\mathbf{z})\,d\mathbf{z}\,,$$

Lipschitz property and convexity follows immediately from those on $\psi(\cdot)$. Moreover, since g_{σ_1} is a probability distribution, we always have $\max|\psi_{\sigma_1}(\mathbf{u})| \le \max|\psi(\mathbf{u})|$. Taking derivatives, we have if $\psi(\cdot)$ is monotonic,

$$\frac{\partial\psi_{\sigma_1}(\mathbf{u})}{\partial u_i} = \nabla\psi_{\sigma_1}(\mathbf{u})^\top\mathbf{e}_i = \int_{\mathbb{R}^K}\nabla\psi(\mathbf{z})^\top\mathbf{e}_i\cdot g_{\sigma_1}(\mathbf{u}-\mathbf{z})\,d\mathbf{z} > 0$$

therefore $\psi_{\sigma_1}(\cdot)$ is also monotonic. Moreover, from Lemma 12 we know $\psi_{\sigma_1}(\cdot)$ is smooth with parameter $\beta = \frac{\sqrt{K}L}{\sigma_1}$ and is L-Lipschitz, and that the mean-squared-error in gradient estimate $\hat{\mathbf{g}}$ is bounded by $\kappa = \tilde{\mathcal{O}}\Big(\frac{L^{7/4}K^{13/8}}{m\sigma_1^{1/4}} + \frac{LK^{5/2}\bar{\epsilon}^2}{\sigma_1}\Big)$ from Lemma 11. Applying Theorem 2 on the smoothed metric $\psi_{\sigma_1}(\cdot)$ with $\eta = \frac{1}{\beta^2} = \frac{\sigma_1^2}{KL^2}$ then completes the proof. □

C. Surrogate PGD as Optimizing a Linear Combination of Surrogates

In this section, we provide an interpretation of Algorithm 1 as optimizing an *adaptively chosen* linear combination of the surrogates $\boldsymbol{\ell}(\theta)$ with an additional proximal penalty like term. Recall that Step 6 of the surrogate projected gradient descent algorithm in Algorithm 1 solves the following optimization problem:

$$\theta^{t+1} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d}\|\big(\boldsymbol{\ell}(\theta) - \tilde{\mathbf{u}}^{t+1}\big)_+\|^2. \tag{23}$$

Lemma 13. *The optimization problem in* (23) *is equivalent to:*

$$\theta^{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\langle \hat{\mathbf{g}}^t, \boldsymbol{\ell}(\theta) \right\rangle + \mathbb{D}(\theta, \theta^t),$$

where $\mathbb{D}(\theta, \theta^t) = \frac{1}{2\eta} \left\| \boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) \right\|^2 + \frac{1}{2\eta} \left\| (\boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t)_+ \right\|^2 - \frac{1}{2\eta} \left\| (\boldsymbol{\ell}(\theta^t) - \eta\, \hat{\mathbf{g}}^t - \boldsymbol{\ell}(\theta)\,)_+ \right\|^2$.

Thus (23) can be seen as minimizing a sum of linear combination of the surrogates and (roughly speaking) a term penalizing some form of distance between the current iterate θ^{t+1} and the previous iterate θ^t.

Proof. Expanding the optimization problem in (23):

$$\theta^{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \| \left(\boldsymbol{\ell}(\theta) - (\boldsymbol{\ell}(\theta^t) - \eta\, \hat{\mathbf{g}}^t) \right)_+ \|^2.$$

Using the identity $(x)_+ = \frac{x + |x|}{2}$, we can write the objective in the above problem as

$$\begin{aligned} &\frac{1}{4} \left\| \boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t + |\boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t| \right\|^2 \\ &\quad = \frac{1}{2} \left\| \boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t \right\|^2 + \frac{1}{2} \left\langle \boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t, |\boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t| \right\rangle \end{aligned}$$

which by ignoring constant terms and noticing that the second term is positive for the coordinates for which $\ell_k(\theta) > \ell_k(\theta^t) - \eta\, \hat{\mathbf{g}}^t_k$ and negative otherwise, we have that

$$\theta^{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\langle \hat{\mathbf{g}}^t, \boldsymbol{\ell}(\theta) \right\rangle + \frac{1}{2\eta} \left\| \boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) \right\|^2 + \frac{1}{2\eta} \left\| (\boldsymbol{\ell}(\theta) - \boldsymbol{\ell}(\theta^t) + \eta\, \hat{\mathbf{g}}^t)_+ \right\|^2 - \frac{1}{2\eta} \left\| (\boldsymbol{\ell}(\theta^t) - \eta\, \hat{\mathbf{g}}^t - \boldsymbol{\ell}(\theta)\,)_+ \right\|^2,$$

as desired. □

D. Additional Experimental Details

D.1. Choice of Hyper-parameters

For the inner projection step in Algorithm 1, we run Adagrad with a fixed step-size of 1.0 for 100 iterations. We used Adagrad as the optimization method for each of the baselines (including logistic regression, and the Relaxed F-measure approach and the Generalized Rates approach in Section 6.2). We tuned the hyper-parameters such as the step size η for the proposed surrogate PGD algorithm and for the baseline Adagrad solvers, and the perturbation parameter σ for gradient estimation in Algorithm 3 using a held-out validation set.

For the F-measure experiments in Section 6.2, we chose the step sizes from the range $\{0.05, 0.1, 0.5, 1.0, 5.0\}$ and σ from the range $\{0.05, 0.1, 0.5\}$. For the ranking experiments in Section 6.3, we chose the step sizes from $\{0.001, 0.005, 0.01\}$ and found a fixed σ of 1.5 to work well across all runs. For the proxy label experiments in Section 6.4, we chose the step sizes from the range $\{0.01, 0.05, 0.1, 0.5, 1.0\}$ and σ from the range $\{0.01, 0.05, 0.1, 0.5, 1.0\}$. For the label noise experiments in Section 6.5, we find a step size of 0.1 and perturbation parameter σ of 1.0 to work well across experiments. For this experiment, we run the projected gradient descent with 300 outer iterations and 100 perturbations.

We implement the metric-optimized example weights approach of Zhao et al. (2019b) in Section 6.5 using the exhaustive search strategy prescribed in their paper. MOEW optimizes a black-box metric by learning a weighted training objective, where the weights on the individual examples are trained to minimize a given metric on the validation set. For a training example (x, y), we compute the weights as a linear function of a 2-dimensional feature embedding $g(x) \in \mathbb{R}^2$ and the labels y, i.e. $w(x, y) = \beta_1 g_1(x) + \beta_2 g_2(x) + \beta_3 y + \beta_4$, and tune the parameters $\beta \in \mathbb{R}^4$ using an exhaustive search over the 4-dimensional grid $\{1/9, \ldots, 8/9\}^4$. The lower-dimensional feature embedding $g(x)$ is computed with principal components analysis. For each choice of candidate weighting function, we train a linear model by minimizing the resulting weighted training objective with 500 steps of Adagrad with step size 0.1, and among the 4096 trained models, pick the one with the least G-mean on the validation set.

Table 7. Additional label noise experiment. Test F-measure on simulated dataset with noisy training labels, averaged over 5 trials. The proposed method was run with sigmoid surrogates. *Higher* is better.

	LogReg	PostShift	MOEW	Proposed
Simulated	0.000	0.172	0.244	**0.287**

Table 8. Average test macro F-measure across groups with clean features. *Higher* is better. We compare the results for the proposed method with 10 and 1000 perturbations to estimate gradients.

	#perturbations = 10	#perturbations = 1000
Business	0.796	0.796
COMPAS	0.630	0.629
Adult	0.661	0.665
Default	0.532	0.533

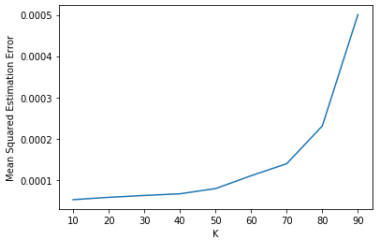

Figure 5. Mean squared estimation error for gradients estimated by the local linear interpolation approach in Algorithm 3 for a synthetic K-dimensional gradient estimation problem, as K varies.

D.2. Additional Label Noise Experiment on Simulated Data

We include an additional experiment for the classification with label noise setting in Section 6.5. We use the simulated data in Section 6.1, and flip a randomly chosen 30% of the positive labels in the training set to negative, and use a clean validation set of size 100. We seek to maximize the F-measure metric. We train linear models and report the test F-measure averaged over 5 trials in Table 7. We again compare with the MOEW approach of Zhao et al. (2019b) and implement it with an exhaustive grid search to tune the weighting function parameters. For this experiment, we directly use the two training features to compute the weighting function instead of a lower-dimensional embedding of the features. The proposed approach is able to adapt better to the noise in the training set and outperforms the other methods.

D.3. Choice of Number of Perturbations

In our experiments in Sections 6.1–6.4, we chose to use 1000 perturbations to estimate gradients as this was a sufficiently large number that worked well for all experiments. But for many experiments, we could get comparable results with fewer perturbations. For example for the experiments in Section 6.1, with as few as 10 perturbations, our approach achieved a test G-mean of 0.801, a comparable value to what we report in Table 2 for the proposed method (0.803). Similarly, for the macro F-measure experiments in Table 3, we got comparable results with 10 perturbations, as shown in Table 8. For the larger KDD Cup 2008 dataset in the ranking experiments in Section 6.3, we used minibatches of size 100 and only perturb examples within each batch to estimate the gradients.

D.4. Dependence of the Gradient Estimation Error on K

While the error bound for the linear interpolation based gradient estimation approach in Lemma 4 has a strong dependence on the number of surrogates K, we find that in our simulations, than the dependence to is less severe. This is evident from the plot shown in Figure 5, where we consider the toy problem of estimating the gradient of the function $f(z) = \left(\prod_{k=1}^{K} z_k\right)^{1/K}$, where $z \in \mathbb{R}_+^K$, and we draw each coordinate z_k from $0.1 + \text{Unif}(0, 0.9)$, We use the local linear interpolation based approach in Algorithm 3 to estimate gradients for f and evaluate the mean squared error for the gradient estimates w.r.t. the true gradient of f as K varies. We use 100 perturbations, and report the average estimation errors over 100 random draws of z and over 100 random trials for each draw of z.