# Supplementary Materials
# StochasticRank: Global Optimization of Scale-Free Discrete Functions

**Aleksei Ustimenko** [1] **Liudmila Prokhorenkova** [1 2 3]

*Table 1.* Notation.

| Variable | Description |
|---|---|
| $z \in \mathbb{R}^n$ | Vector of scores |
| $\xi \in \Xi_n$ | Vector of contexts |
| $r \in \mathbb{R}^n$ | Vector of relevance labels |
| $\theta \in \mathbb{R}^m$ | Vector of parameters |
| $L(z, \xi)$ | Loss function |
| $L_\xi^\pi(z, \sigma)$ | Smoothed loss function |
| $L_\xi^\pi(z, \sigma\|z')$ | SFA smoothing of the loss |
| $\mathcal{L}(\theta)$ | Expected loss |
| $\mathcal{L}_N(\theta)$ | Empirical loss |
| $\mathcal{L}_N^\pi(\theta, \sigma)$ | Smoothed empirical loss |
| $\mathcal{L}_N^\pi(\theta, \sigma, \gamma)$ | Regularized and consistently smoothed loss |
| $\mathcal{R}_0$ | Scale-free discrete loss functions |
| $\mathcal{R}_1$ | Ranking loss functions |
| $\mathcal{R}_1^{soft}$ | Soft ranking loss functions |
| $\pi_\xi(z)$ | Distribution density for smoothing |
| $p_\beta(\theta)$ | Invariant measure of parameters |
| $p_\beta(F)$ | Invariant measure of predictions |
| $\sigma > 0$ | Smoothing standart deviation |
| $\beta > 0$ | Diffusion temperature |
| $\gamma > 0$ | Regularization parameter |
| $\mu \geq 0$ | Relevance shifting parameter |
| $\nu > 0$ | Scale-Free Acceleration parameter |

## A. Proof of Statement 1

Let us prove that the set $\arg\min_{\theta \in \mathbb{R}^m} \mathcal{L}_N(\theta)$ is not empty.

Consider $U_{ij}$ being open and convex sets for $V_i = \mathrm{im}\Phi_{\xi_i}$ (see Discreteness on subspaces in Definition 1 in the main text). Then, $U'_{ij} = \Phi_{\xi_i}^{-1} U_{ij} \subset \mathbb{R}^m$ are also open and convex. Henceforth, the function $\mathcal{L}_N$ can be written as (ignoring the sets of zero measure):

$$\mathcal{L}_N(\theta) = N^{-1} \sum_{j_1=1}^{k_1} \dots \sum_{j_N=1}^{k_N} c_{j_1,\dots j_N} \mathbb{1}_{\theta \in \cap_{i=1}^N U'_{ij_i}} \,. \tag{1}$$

Henceforth, the function $\mathcal{L}_N$ is also discrete with open convex sets $\mathcal{U}_s := \cap_{i=1}^N U'_{ij_i}$ on the whole space $\mathbb{R}^m$. Hence, its $\arg\min$ is one of these sets or their union.

## B. Stochastic smoothing

### B.1. Mollification

A natural approach for smoothing is mollification (Ermoliev et al., 1995; Dolecki et al., 1983): choose a smooth enough distribution with p.d.f. $\pi(\theta)$, consider the family of distributions $\pi_\delta(\theta) = \delta^{-m}\pi(\delta^{-1}\theta)$, and let $\mathcal{L}_N(\theta, \delta) := \mathcal{L}_N * \pi_\delta \equiv \mathbb{E}_{\epsilon\sim\pi}\mathcal{L}_N(\theta + \delta\epsilon)$. Then, the minimizers of $\mathcal{L}_N(\theta, \delta)$ convergence to the minimizer of $\mathcal{L}_N(\theta)$. Unfortunately, despite theoretical soundness, it is hard to derive efficient gradient estimates even in the linear case $f_{\xi_i}(\theta) = \Phi_{\xi_i}\theta$. Moreover, in the gradient boosting setting, we do not have access to all possible coordinates of $\theta$ at each iteration. Henceforth, we cannot use the mollification approach directly.

Thus, instead of acting on the level of parameters $\theta$, we act on the level of scores $z$: $L_\xi^\pi(z, \sigma) := \mathbb{E}L(z + \sigma\varepsilon, \xi)$, where $\varepsilon$ has p.d.f. $\pi(z)$. We multiply the noise by $\sigma$ to preserve Scalar-freeness in a sense that $L_\xi^\pi(\lambda z, \lambda\sigma) = L_\xi^\pi(z, \sigma)$ for any $\lambda > 0$.

In the linear case $f(\theta) = \Phi\theta$, if $\mathrm{rk}\Phi = n$, it is not hard to show the convergence of minimizers. Indeed, we can obtain mollification by "bypassing" the noise from scores to parameters by multiplying on $\Phi^{-1}$. However, in general, we cannot assume $\mathrm{rk}\Phi = n$.

### B.2. Proof of Theorem 1

The trick is to proceed with $L(f_{\xi_i}(\theta), \xi_i)$ and to show that there exists an open and dense set $U_{\xi_i} \subset \mathbb{R}^m$ such that the convergence is locally uniform as $\sigma \to 0_+$, $\mu \to \infty$, $\sigma\mu \to 0_+$.

Let us proceed with proving the existence of such $U_{\xi_i} \forall i$. Let us define

$$U_{\xi_i} := \Big\{\theta \in \mathbb{R}^m : \forall j \neq j' \big(f_{\xi_i}(\theta)_j = f_{\xi_i}(\theta)_{j'}\big) \Rightarrow \\ \forall \theta' \in \mathbb{R}^m \big(f_{\xi_i}(\theta')_j = f_{\xi_i}(\theta')_{j'}\big)\Big\}.$$

[1]Yandex, Moscow, Russia [2]Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia [3]Higher School of Economics, Moscow, Russia. Correspondence to: Aleksei Ustimenko <austimenko@yandex-team.ru>.

*Proceedings of the 37<sup>th</sup> International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Clearly, the set is not empty, open, and dense. Now, take an arbitrary $\theta \in U_{\xi_i}$. Consider $z = f_{\xi_i}(\theta)$ and divide the set $\{1, \ldots, n_i\}$ into disjoint subsets $J_1, \ldots, J_k$ such that all components $z_j$ corresponding to one group are equal and all components $z_j$ corresponding to different $J$'s are different. Clearly, we need to "resolve" only those which are equal: for small enough $\sigma \approx 0, \sigma\mu \approx 0$ we obtain that even after adding the noise $f_{\xi_i}(\theta') - \sigma\mu r + \sigma\varepsilon$ the order of $J$'s is preserved with high probability uniformly in some vicinity of $\theta$, whilst for large enough $\mu \gg 1$ we obtain the worst case permutation of $z_j$ corresponding to the one group with high probability uniformly on the whole $U_{\xi_i}$. Thus, we obtain locally uniform convergence $\mathbb{E}L(f_{\xi_i}(\theta) - \sigma\mu r + \sigma\varepsilon, \xi_i) \to L(f_{\xi_i}(\theta), \xi_i)$.

### B.3. Proof of Theorem 2

Clearly, the conditions of the theorem imply that for general $\theta$ w.l.o.g. we can assume that $\Phi_{\xi_i}\theta \in U_{ij_i}$ for some indexes $j_i$. Henceforth, after adding the noise with $\sigma \to 0_+$ we must obtain locally uniform approximation since the functions $L(z, \xi_i)$ are locally constant in a vicinity of $z = \Phi_{\xi_i}\theta \, \forall i$.

### B.4. Consistent smoothing for LSO

**Theorem 1.** *In gradient boosting, if $L(\cdot,\cdot) \in \mathcal{R}_0$ is coming from the LSO problem, then any smoothing is consistent.*

*Proof.* Conditions from Theorem 2 of the main text translate into a condition that $\left(\Phi_\xi\theta\right)_j \neq 0$ for all $j$ and for all $\theta$ almost surely. This can be enforced by adding a free constant to the linear model, but in the gradient boosting setting this condition is essentially satisfied: consider $\theta = \mathbb{1}_m$, then $\left(\Phi_\xi \mathbb{1}_m\right)_j \geq 1 \, \forall j$ since the matrix $\Phi_\xi$ is 0-1 matrix and have at least one "1" in each row (every item fells to at least one leaf of each tree). Henceforth, for any general $\theta$ we can assume another general $\widetilde{\theta} = \theta + \nu\mathbb{1}_m$, where $\nu$ is any random variable with absolute continuous p.d.f. This in turn implies $\left(\Phi_\xi\widetilde{\theta}\right)_j \neq 0$ almost surely. Henceforth, Theorem 2 holds ensuring the consistency of smoothing. $\square$

## C. Coordinate Conditional Sampling

### C.1. Proof of Lemma 1

Consider a line $H = \{(z_j, z_{\setminus j}) : \forall z_j \in \mathbb{R}\}$ and subsets $U_1, \cdots, U_k$ for $k = k(n, \mathbb{R}^n)$ from the Discretness on subspaces assumption for $V = \mathbb{R}^n$. Then $U_i \cap H = (a_i, b_i) \times \{z_{\setminus j}\}$ due to opennes and convexity of $U_i$ for $a_i, b_i \in \mathbb{R} \cup \{\pm\infty\}$. Moreover, $(U_i \cap H) \cap (U_{i'} \cap H) = \emptyset \, \forall i \neq i'$ and, by ignoring sets of zero measure, we can assume that $\overline{\cup_i (a_i, b_i) \times \{z_{\setminus j}\}} = H$. After that, we can take all finite $\{b_1, \ldots, b_k\} \cap \mathbb{R}$ as breaking points.

### C.2. Proof of Theorem 3

Observe that $L * \pi_\xi^j$ tautologically equals $l_j * \pi_\xi^j$ and the convolution is distributive with respect to summation, so we can write:

$$L * \pi^j = \sum_{s=1}^{k'} \Delta l_j(b_s) \mathbb{1}_{\{z_j \leq b_s\}} * \pi_\xi^j + \mathrm{const}(z_{\setminus j}).$$

The convolution $\mathbb{1}_{\{z_j \leq b_s\}} * \pi_\xi^j$ is equal to $\mathbb{P}_\xi(z_j + \sigma\varepsilon_j < b_s | \varepsilon_{\setminus j}) := \sigma^{-1} \int_{\mathbb{R}} \mathbb{1}_{\{z_j + \sigma\varepsilon_j \leq b_s\}} \pi_\xi^j(\sigma^{-1}\varepsilon_j) \mathrm{d}\varepsilon_j$, allowing us to rewrite:

$$\begin{aligned} &L * \pi_\xi^j \\ &= \sum_{s=1}^{k'} \Delta l_j(b_s) \mathbb{P}_\xi(\varepsilon_j < \sigma^{-1}(b_s - z_j) | \varepsilon_{\setminus j}) + \mathrm{const}(z_{\setminus j}). \end{aligned}$$

The above formula is ready for differentiation since each term is actually a $C^{(2)}(\mathbb{R})$ function by the variable $z_j$:

$$\frac{\partial}{\partial z_j} L * \pi_\xi^j = -\sigma^{-1} \sum_{s=1}^{k'} \Delta l_j(b_s) \pi^j(\sigma^{-1}(b_s - z_j)).$$

After the convolution with $\pi_\xi^{\setminus j}$, we finally get the required formula.

### C.3. Proof of Corollary 1

For LTR ($\mathcal{R}_1$ and $\mathcal{R}_1^{soft}$), all these $b_s$ actually lay in $\{z_1, \ldots, z_n\} \subset \mathbb{R}$ due to Pairwise decision boundary assumption and, henceforth, we do not need to compute them, we just need to take coordinates of $z \in \mathbb{R}^n$ as breaking points and note that if some of $z_s$ is not a breaking point for $L(z, \xi)$, then essentially $\Delta l_j(z_s) = 0$. Then, we can write

$$\frac{\partial}{\partial z_j} L * \pi_\xi^j = -\sigma^{-1} \sum_{s=1}^{n} \Delta l_j(z_s) \pi_\xi^j(\sigma^{-1}(z_s - z_j)).$$

Let us note that for LSO, we can actually take $k' = 1$ and $b_1 = 0$ and simplify the formula to:

$$l_j(z_j) = \Delta l_j \mathbb{1}_{\{z_j \leq 0\}} + \mathrm{const}(z_{\setminus j}).$$

### C.4. Proof of Theorem 4

**Lemma 1.** *The function $L_\xi^\pi(z, \sigma)$ satisfies the following linear first order Partial Differential Equation (PDE):*

$$\frac{\partial}{\partial \sigma} L_\xi^\pi(z, \sigma) = -\sigma^{-1} \langle \nabla_z L_\xi^\pi(z, \sigma), z \rangle_2.$$

*Proof.* The proof is a direct consequence of Scalar-Freenees: we just need to differentiate the equality $L_\xi^\pi(\alpha z, \alpha\sigma) \equiv L_\xi^\pi(z, \sigma)$ (holding for $\alpha > 0$) by $\alpha$ and set $\alpha = 1$. $\square$

**Lemma 2.** $\frac{\partial}{\partial\sigma} L^{\pi}_{\xi}(z,\sigma)$ *is uniformly bounded by* $\mathcal{O}(\sigma^{-1})$.

*Proof.* Consider writing $L^{\pi}_{\xi}(z,\sigma)$ in the integral form:

$$L^{\pi}_{\xi}(z,\sigma) = \sigma^{-n} \int_{\mathbb{R}^n} L(z+\varepsilon,\xi)\pi(\sigma^{-1}\varepsilon)\mathrm{d}\varepsilon.$$

By Fubini's theorem, we can pass the differentiation $\frac{\partial}{\partial\sigma}$ to inside the integral and obtain:

$$\frac{\partial}{\partial\sigma} L^{\pi}_{\xi}(z,\sigma) = -n\sigma^{-n-1} \int_{\mathbb{R}^n} L(z+\varepsilon,\xi)\pi(\sigma^{-1}\varepsilon)\mathrm{d}\varepsilon - \sigma^{-n-2} \int_{\mathbb{R}^n} L(z+\varepsilon,\xi)\langle\nabla\pi(\sigma^{-1}\varepsilon),\varepsilon\rangle\mathrm{d}\varepsilon.$$

Consider the variable $\varepsilon' = \sigma^{-1}\varepsilon$, then we arrive at

$$\frac{\partial}{\partial\sigma} L^{\pi}_{\xi}(z,\sigma) = -n\sigma^{-1} \int_{\mathbb{R}^n} L(z+\sigma\varepsilon,\xi)\pi(\varepsilon)\mathrm{d}\varepsilon - \sigma^{-1} \int_{\mathbb{R}^n} L(z+\sigma\varepsilon,\xi)\langle\nabla\pi(\varepsilon),\varepsilon\rangle\mathrm{d}\varepsilon.$$

Taking the absolute value of both sides and using the triangle inequality, we derive

$$\left|\frac{\partial}{\partial\sigma} L^{\pi}_{\xi}\right| \;\leq\; nl\sigma^{-1} \;+\; l\sigma^{-1} \int_{\mathbb{R}^n} \|\nabla\pi(\varepsilon)\|_2 \|\varepsilon\|_2 \mathrm{d}\varepsilon,$$

where $l = \sup_z |L(z,\xi)| < \infty$ by the Uniform boundedness assumption and the last integral is well defined by the Derivative decay assumption. □

**Corollary 1.** $\sup_z \left|\left\langle \nabla_z L^{\pi}_{\xi}, z \right\rangle_2\right| = \mathcal{O}(1)$ *independently from* $\sigma$.

*Proof.* Immediate consequence of the previous lemmas. □

Now, assume that $\sigma = \sigma(z)$ is differentiable and non-zero at $z$. The following lemma describes $\nabla_z L^{\pi}_{\xi}(z,\sigma(z))$ in terms of $\nabla_z L^{\pi}_{\xi} := \nabla_z L^{\pi}_{\xi}(z,\sigma)\big|_{\sigma=\sigma(z)}$.

**Lemma 3.** *The following formula holds:*

$$\nabla_z L^{\pi}_{\xi}(z,\sigma(z)) = \nabla_z L^{\pi}_{\xi} - \left\langle \nabla_z L^{\pi}_{\xi}, z \right\rangle_2 \nabla_z \log\sigma(z).$$

*Proof.* Consider writing

$$\nabla_z L^{\pi}_{\xi}(z,\sigma(z)) = \nabla_z L^{\pi}_{\xi} + \frac{\partial}{\partial\sigma} L^{\pi}_{\xi}(z,\sigma(z)) \nabla_z\sigma(z).$$

Then, by Lemma 1 we obtain the formula. □

## D. Fast ranking metrics computation

We need to be able to compute $L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi)$ for an arbitrary $z' \in \mathbb{R}$ and a position $i$, where $s \in S_n$ represents $s := \mathrm{argsort}(z+\sigma\varepsilon)$ for the CCS estimate (note that there is no ambiguity in computing $\mathrm{argsort}$ since with probability one $z_{j_1} + \sigma\varepsilon_{j_1} \neq z_{j_2} + \sigma\varepsilon_{j_2}$ for $j_1 \neq j_2$). Moreover, $\mathrm{argsort}$ requires $\mathcal{O}(n\log n)$ operations.

Typically, the evaluation of $L(\cdots)$ costs $\mathcal{O}(n)$, e.g., for $\mathrm{ERR}$. Fortunately, for many losses it is possible to exploit the structure of the loss that allows evaluating $L$ in $\mathcal{O}(1)$ operations using some precomputed shared cumulative statistics related to the loss which can be computed in $\mathcal{O}(n)$ operations and $\mathcal{O}(n)$ memory.

For all $L \in \mathcal{R}_1$ in the worst case we need $\mathcal{O}(n^2)$ evaluations of $L$ to compute the CCS (for each of $n$ coordinates to sum up at most $n$ evaluations). Thus, the overall worst case asymptotic of the algorithm would be $\mathcal{O}(n\log n + n + n^2) = \mathcal{O}(n^2)$ if the evaluation costs $\mathcal{O}(1)$. For the sake of simplicity, we generalize both $\mathrm{NDCG@k}$ and $\mathrm{ERR}$ into one class of losses:

$$L(z,\xi) = -\sum_{i=1}^{n} w_i g(r_{s_i}) \prod_{j=1}^{i-1} d_{s_j}, \tag{2}$$

where $W = \{w_i\}_{i=1}^n$ are some predefined positions' weights typically picked as $\frac{\mathbb{1}_{\{i\leq k\}}}{\max_z \mathrm{DCG@k}\log(i+1)}$ for $\mathrm{NDCG@k}$ and $\frac{1}{i}$ for ERR); $D = \{d_i\}_{i=1}^n$ is typically picked as $d_i = 1 \forall i$ for $-\mathrm{NDCG@k}$ and $d_i = 1 - r_i \forall i$ for ERR; and finally we define $g(r) = r$ for $r \in [0,1]$ and $g(r) = \frac{2^r - 1}{2^4}$ for $r \in \{0,1,2,3,4\}$.

First, we need to define and compute the following cumulative product:

$$p_m = d_{s_{m-1}} p_{m-1} = \prod_{j=1}^{m-1} d_{s_j} \text{ if } m > 1,$$

where $p_1 = 1$. Denote $P := \{p_i\}_{i=1}^n$. Next, we use them we define the following cumulative sums:

$$S^{\mathrm{up}}_m = S^{\mathrm{up}}_{m-1} + w_{m+1} g(r_{s_m}) p_m \text{ if } m > 1,$$

$$S^{\mathrm{mid}}_m = S^{\mathrm{mid}}_{m-1} + w_m g(r_{s_m}) p_m \text{ if } m > 0,$$

$$S^{\mathrm{low}}_m = S^{\mathrm{low}}_{m-1} + w_{m-1} g(r_{s_m}) p_m \text{ if } m > 0,$$

where $S^{\mathrm{up}}_0 = S^{\mathrm{up}}_1 = S^{\mathrm{mid}}_0 = S^{\mathrm{low}}_0 = 0$.

All these cumulative statistics can be computed at the same time while we compute $L(z+\sigma\varepsilon,\xi)$. Note that we need additional $O(n)$ memory to store these statistics.

Now fix a position $i$ and score $z'$. Express $L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi)$ as $(L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi) - L(z+\sigma\varepsilon,\xi)) + L(z+$

$\sigma\varepsilon, \xi)$. Thus, we need to compute $L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi) - L(z + \sigma\varepsilon, \xi)$.

If $z' > z_{s_i} + \sigma\varepsilon_{s_i}$, we define $i' := i$; otherwise, define $i' := i - 1$ — this variable represents the new position of the $s_i$-th document in $z + \sigma\varepsilon$. Also, if $z' > z_{s_i} + \sigma\varepsilon_{s_i}$, we define:

$$T^{\text{low}} = S_{i'}^{\text{mid}} - S_i^{\text{mid}},$$
$$T^{\text{up}} = d_{s_i}^{-1}(S_{i'}^{\text{up}} - S_i^{\text{up}}),$$
$$w = w_i p_i,$$
$$w' = w_{i'} d_{s_i}^{-1} p_{i'}.$$

Otherwise, define:

$$T^{\text{low}} = d_{s_i}(S_{i'}^{\text{low}} - S_{i-1}^{\text{low}}),$$
$$T^{\text{up}} = S_{i'}^{\text{mid}} - S_{i-1}^{\text{mid}},$$
$$w = w_i p_i,$$
$$w' = w_{i'-1} p_{i'}.$$

Then, we calculate $L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi) - L(z + \sigma\varepsilon, \xi)$ as $g(r_{s_i})(w - w') - (T^{\text{up}} - T^{\text{low}})$. The meaning of the formula is simple: we measure the change of gain of the $s_i$-th document if we change its score to $z'$ from $z_{s_i} + \sigma\varepsilon_{s_i}$ minus the difference of gains of all documents on positions from $i'$ up to $i - 1$, if $i' < i$, and from $i + 1$ up to $i' - 1$, if $i' > i$.

The above formulas can be verified directly by evaluating the cases when $z' > z_{s_i} + \sigma\varepsilon_{s_i}$ or $z' < z_{s_i} + \sigma\varepsilon_{s_i}$ and expanding $S_m^*$ as $\sum_i w_{i\pm1} g(r_{s_i}) p_i$. Note that all differences $S_i^* - S_j^*$ take into account all documents on positions from $j + 1$ up to $i$ inclusively.

Note that $S_n^{\text{mid}} \equiv L(z + \sigma\varepsilon, \xi)$. Indeed,

$$\sum_{i=1}^{n} w_i g(r_{s_i}) p_i = \sum_{i=1}^{n} w_i g(r_{s_i}) \prod_{j=1}^{i-1} d_{s_j} = L(z + \sigma\varepsilon, \xi).$$

Therefore, we obtain:

$$L(z', z_{\setminus s_i} + \sigma\varepsilon_{\setminus s_i}, \xi) = g(r_{s_i})(w - w') - (T^{\text{up}} - T^{\text{low}}) + S_k^{\text{mid}}. \quad (3)$$

# E. Global Optimization by Diffusion

## E.1. Overview of SGLB idea

Global convergence of SGLB is guaranteed by a so-called Predictions' Space Langevin Dynamics Stochastic Differential Equation

$$\mathrm{d}F(t) = -\gamma F(t)\mathrm{d}t - P\nabla_F \mathcal{L}_N^{\pi}(F(t), \sigma)\mathrm{d}t + \sqrt{2\beta^{-1}P}\mathrm{d}W(t),$$

where $F(t) := \Phi\theta(t) = (\Phi_{\xi_1}\theta(t), \dots, \Phi_{\xi_N}\theta(t)) = (f_{\xi_1}(\theta), \dots, f_{\xi_N}(\theta)) \in \mathbb{R}^{N'}$ denotes the predictions Markov Process on the train set $\mathcal{D}_N$, $W(t)$ is a standard Wiener process with values in $\mathbb{R}^{N'}$, $N' := \sum_{i=1}^{N} n_i$, $P = P^T$ is an implicit preconditioner matrix of the boosting algorithm, and $\beta > 0$ is a temperature parameter that controls exploration/exploitation trade-off. Note that here we override the notation $\mathcal{L}_N(F) \equiv \mathcal{L}_N(\theta)$ since $F = \Phi\theta$. Further by $\Gamma = \sqrt{P^{-1}}$ we denote an implicitly defined regularization matrix.

The global convergence is implied by the fact that as $t \to \infty$, the stationary distribution $p_\beta(F)$ of $F(t)$ concentrates around the global optima of the implicitly regularized loss

$$\mathcal{L}_N^{\pi}(F, \sigma, \gamma) = \mathcal{L}_N^{\pi}(F, \sigma) + \frac{\gamma}{2}\|\Gamma F\|_2^2 .$$

More formally, the stationary distribution is $p_\beta(F) \propto \exp(-\beta\mathcal{L}_N^{\pi}(F, \sigma, \gamma))$. According to Ustimenko & Prokhorenkova (2020), optimization is performed within a linear space $V := \operatorname{im}\Phi$ that encodes all possible predictions $F$ of all possible ensembles formed by the weak learners associated with the boosting algorithm. We refer interested readers to (Ustimenko & Prokhorenkova, 2020) for the details.

## E.2. Proof of Theorem 5

Let us first prove the following lemma.

**Lemma 4.** *The function $\mathcal{L}_N^{\pi}(F, \sigma)$ is uniformly bounded, Lipschitz continuous with constant $L_0 = \mathcal{O}(\sigma^{-1})$, and Lipschitz smooth with constant $L_1 = \mathcal{O}(\sigma^{-2})$.*

*Proof.* The proof of Lipschitz continuity is a direct consequence of the uniform boundedness by $\mathcal{O}(\sigma^{-1})$ of CCS. If we differentiate CCS estimate one more time, we obtain the estimates for the Hessian that must be uniformly bounded by $\mathcal{O}(\sigma^{-2})$ due to the uniform boundedness of $\nabla\pi$, thus giving Lipschitz smoothness. □

In addition to Lipschitz smoothness, continuity, and boundedness from above, we also need $\|\widehat{\nabla}_{CC}\mathcal{L}_N^{\pi}(F, \sigma) - \nabla\mathcal{L}_N^{\pi}(F, \sigma)\|_2 = \mathcal{O}(1)$ (Ustimenko & Prokhorenkova, 2020), but that condition is satisfied since both terms are uniformly bounded by $\mathcal{O}(\sigma^{-1})$. Thus, the algorithm has limiting stationary measure $p_\beta(F) \propto \exp(-\beta\mathcal{L}_N^{\pi}(F, \sigma, \gamma))$.

Then, consistency of the smoothing ensures that as $\sigma \to 0_+$, $p_\beta(F) \to p_\beta^*(F)$, where $p_\beta^*(F) \propto \exp(-\beta(\mathcal{L}_N(F) + \frac{\gamma}{2}\|\Gamma F\|_2^2))$ and thus for $\beta \gg 1$ the measures $p_\beta^*$ and $p_\beta$ for $\sigma \approx 0$ concentrate around the global optima of $\mathcal{L}_N(F)$.

## E.3. Proof of Theorem 6

Following Raginsky et al. (2017); Ustimenko & Prokhorenkova (2020), we immediately obtain that

$\left|\mathbb{E}_{\theta\sim p_\beta(\theta)}\mathcal{L}^\pi(\theta,\sigma) - \mathbb{E}_{\theta\sim p_\beta(\theta)}\mathcal{L}^\pi_N(\theta,\sigma)\right| = \mathcal{O}(\frac{(\beta+d)^2}{N\lambda_*})$ with $\lambda_* > 0$ and $d = V_\mathcal{B}$. In general non-convex case $\frac{1}{\lambda_*}$ can be of order $\exp(\mathcal{O}(d))$ (Raginsky et al., 2017) but for smoothed SF losses we can give a better estimate without exponential dependence on the dimension.

Observe that our measure is the sum of uniformly bounded Lipschitz smooth with constant $\mathcal{O}(\sigma^{-2})$ and a Gaussian $\frac{\gamma}{2}\|\Gamma\Phi\theta\|_2^2$, then the more appropriate bound from the logarithmic Sobolev inequality applies according to Lemma 2.1 (Bardet et al., 2015) $\frac{1}{\lambda_*} = \mathcal{O}\left(\frac{\exp(\mathcal{O}(\frac{\beta}{\gamma\sigma^2}))}{\gamma\beta}\right)$ being dimension-free. Note that Miclo's trick in the proof of the lemma should be skipped since $\mathcal{L}^\pi_N(\theta,\sigma)$ is already fine enough. Coupling the spectral gap bound with the generalization gap, we obtain the theorem.

## F. Parameter tuning

For tuning, we use the random search (500 samples) with the following distributions:

- For *learning-rate* log-uniform distribution over $[10^{-3}, 1]$.
- For *l2-leaf-reg* log-uniform distribution over $[10^{-1}, 10^1]$ for baselines and *l2-leaf-reg=0* for StochasticRank.
- For noise strength (Bruch et al., 2020) uniform distribution over $[0, 1]$.
- For *depth* uniform distribution over $\{6, 7, 8, 9, 10\}$.
- For *model-shrink-rate* log-uniform distribution over $[10^{-5}, 10^{-2}]$ for StochasticRank.
- For *diffusion-temperature* log-uniform distribution over $[10^8, 10^{11}]$ for StochasticRank.
- For *mu* log-uniform distribution over $[10^{-2}, 10]$ for StochasticRank-$\mathcal{R}_1$.

## References


Bardet, J.-B., Gozlan, N., Malrieu, F., and Zitt, P.-A. Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence. *arXiv e-prints*, art. arXiv:1507.02389, 2015.

Bruch, S., Han, S., Bendersky, M., and Najork, M. A stochastic treatment of learning to rank scoring functions. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*, pp. 61–69, 2020.

Dolecki, S., Salinetti, G., and Wets, R. J.-B. Convergence of functions: equi-semicontinuity. *Transactions of the American Mathematical Society*, 276(1):409–429, 1983.

Ermoliev, Y., Norkin, V., and Wets, R. The minimization of semicontinuous functions: mollifier subgradients. *SIAM Journal on Control and Optimization*, 33, 01 1995.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *CoRR*, abs/1702.03849, 2017.

Ustimenko, A. and Prokhorenkova, L. SGLB: Stochastic Gradient Langevin Boosting. *arXiv e-prints*, art. arXiv:2001.07248, 2020.