PRECISION SAMPLING

Yanheng Wang

March 16, 2024

Precision sampling, proposed by Andoni, Krauthgamer and Onak, is a method that estimates a sum using noisy observations of its summands. It can be formulated as a game. Alice keeps private values $a_1, \ldots, a_n \in [0,1]$ and Bob wants to estimate their sum $\sigma := \sum_{i=1}^n a_i$. He tells Alice a tolerable deviation $d_i \geqslant 0$ for each $i \in [n]$, and she in response gives him noisy observations $\hat{a}_i = a_i \pm d_i$. Based on these, Bob should output an estimate $\hat{\sigma}$ such that $\frac{1-\varepsilon}{1+\varepsilon}\sigma - \delta \leqslant \hat{\sigma} \leqslant \frac{1+\varepsilon}{1-\varepsilon}\sigma + \delta$, where $\varepsilon, \delta \geqslant 0$ are prescribed error parameters.

Of course, if Bob requests $d_i \leqslant \delta/n$ for all $i \in [n]$ then he can just sum up the observations to get a good estimate. But this strategy takes heavy toll on Alice's side, for generally, the cost of an observation inversely relates to the deviation d_i. Think Alice as a measuring instrument: the higher precision we demand, the more resource she needs. Now Bob wonders: Is there a better strategy that saves Alice's cost?

Precision Sampling Lemma. Fix $1/2 \geqslant \varepsilon > 0$ and $\delta \geqslant 0$ and write $\ell := 10/\varepsilon^2\delta$. Consider the following strategy of Bob.

```
fn errors()
    for i = 1, ..., n do
        sample independent u_i1, ..., u_iℓ ~ Uniform(0,1)
        let d_i := min{u_i1, ..., u_iℓ}
    send d_1, ..., d_n to Alice

fn sum(â_1, ..., â_n)
    for i = 1, ..., n do
        let s_i count the number of j ∈ [ℓ] : u_ij ⩽ ε â_i
    output σ̂ := (1/(εℓ)) · Σ_{i=1}^n s_i
```

Then with probability at least $3/5$ (regardless of the concrete values returned by Alice), Bob outputs $\frac{1-\varepsilon}{1+\varepsilon}\sigma - \delta \leqslant \hat{\sigma} \leqslant \frac{1+\varepsilon}{1-\varepsilon}\sigma + \delta$. Moreover, for every $i \in [n]$,

$$\mathbb{E}(d_i^{-\theta}) = \begin{cases} O(\ell \log n) & \theta = 1, \\ O\left(\frac{\ell^\theta}{1-\theta}\right) & 0 \leqslant \theta < 1. \end{cases}$$

Proof. The "moreover" part is a routine calculation; just use the fact that d_i has density function $f(x) = \ell(1-x)^{\ell-1}$.

For the remaining part, our plan is to show that $\mathbb{E}(\hat{\sigma}) = \frac{1}{\varepsilon \ell} \sum_{i=1}^n \mathbb{E}(s_i)$ is roughly σ, and that $\hat{\sigma}$ concentrates around its mean. However, since the $\hat{a}_i$'s may adversarially depend on d_i's, it is difficult to analyse $\mathbb{E}(s_i)$ directly. The trick is to sandwich the variables by easier ones. For $i \in [n]$ and $j \in [\ell]$ we define

$$\begin{aligned} X_{ij} &:= \mathbf{1}\left\{u_{ij} \leqslant \frac{\varepsilon a_i}{1+\varepsilon}\right\}, \\ S_{ij} &:= \mathbf{1}\{u_{ij} \leqslant \varepsilon \hat{a}_i\}, \\ Y_{ij} &:= \mathbf{1}\left\{u_{ij} \leqslant \frac{\varepsilon a_i}{1-\varepsilon}\right\}. \end{aligned}$$

Observe that $X_{ij} \leqslant S_{ij} \leqslant Y_{ij}$. Indeed, if $X_{ij} = 1$ then $(1+\varepsilon) u_{ij} \leqslant \varepsilon a_i$, thus

$$u_{ij} \leqslant \varepsilon (a_i - u_{ij}) \leqslant \varepsilon (a_i - d_i) \leqslant \varepsilon \hat{a}_i$$

which means $S_{ij} = 1$. Similarly, if $S_{ij} = 1$ then

$$u_{ij} \leqslant \varepsilon \hat{a}_i \leqslant \varepsilon (a_i + d_i) \leqslant \varepsilon (a_i + u_{ij}),$$

so $u_{ij} \leqslant \varepsilon a_i / (1-\varepsilon)$ and thus $Y_{ij} = 1$. It immediately follows that

$$\left(X := \frac{\sum_{i=1}^n \sum_{j=1}^\ell X_{ij}}{\varepsilon \ell}\right) \leqslant \hat{\sigma} \leqslant \left(Y := \frac{\sum_{i=1}^n \sum_{j=1}^\ell Y_{ij}}{\varepsilon \ell}\right)$$

Now we turn to study the behaviours of X and Y. We calculate

$$\begin{aligned} \mathbb{E}(X) &= \frac{\sum_{i=1}^n \sum_{j=1}^\ell \varepsilon a_i / (1+\varepsilon)}{\varepsilon \ell} = \frac{\sigma}{1+\varepsilon}, \\ \mathbb{E}(Y) &= \frac{\sum_{i=1}^n \sum_{j=1}^\ell \varepsilon a_i / (1-\varepsilon)}{\varepsilon \ell} = \frac{\sigma}{1-\varepsilon}. \end{aligned}$$

Similarly, using independence,

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{\ell} \mathrm{Var}(X_{ij})}{(\varepsilon\,\ell)^2} \leqslant \frac{\mathbb{E}(X)}{\varepsilon\,\ell} = \frac{\varepsilon\,\delta\,\mathbb{E}(X)}{10},$$
$$\mathrm{Var}(Y) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{\ell} \mathrm{Var}(Y_{ij})}{(\varepsilon\,\ell)^2} \leqslant \frac{\mathbb{E}(Y)}{\varepsilon\,\ell} = \frac{\varepsilon\,\delta\,\mathbb{E}(Y)}{10}.$$

To proceed, we distinguish two cases. If the sum is large, specifically $\sigma \geqslant \delta/\varepsilon$, then we can bound the probability that X, Y deviate from σ multiplicatively:

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant \varepsilon\,\mathbb{E}X) \leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2\,\mathbb{E}^2(X)} \leqslant \frac{\delta}{10\,\varepsilon\,\mathbb{E}(X)} \leqslant \frac{\delta}{5\,\varepsilon\,\sigma} \leqslant \frac{1}{5},$$

similar for Y. Therefore, with probability at least $3/5$ both X and Y are within $(1 \pm \varepsilon)$ times their respective expectations. In such event, we have

$$\frac{1-\varepsilon}{1+\varepsilon}\sigma \leqslant \hat{\sigma} \leqslant \frac{1+\varepsilon}{1-\varepsilon}\sigma$$

as desired.

It remains to consider the case when the sum is small: $\sigma < \delta/\varepsilon$. Due to the small expectations, we can only bound additive deviations:

$$\mathbb{P}(|X - \mathbb{E}X| \geqslant \delta) \leqslant \frac{\mathrm{Var}(X)}{\delta^2} \leqslant \frac{\varepsilon\,\mathbb{E}(X)}{10\,\delta} \leqslant \frac{\varepsilon\,\sigma}{5\,\delta} < \frac{1}{5},$$

similar for Y. Therefore, with probability at least $3/5$ both X and Y are within $\pm\delta$ of their respective expectations. In such event, we have

$$\frac{1}{1+\varepsilon}\sigma - \delta \leqslant \hat{\sigma} \leqslant \frac{1}{1-\varepsilon}\sigma + \delta$$

as desired. □

Extensions. We did not present the lemma in its full generality. A few extensions are possible. First, we may allow observation $\hat{a}_i$ to deviate from truth a_i multiplicatively. That is, we allow Alice returning

$$\frac{a_i}{\gamma} - d \leqslant \hat{a}_i \leqslant \gamma\,a_i + d$$

for some constant $\gamma \geqslant 1$. The lemma still holds except that the output is affected by a γ-factor:

$$\frac{1-\varepsilon}{(1+\varepsilon)\gamma}\sigma - \delta \leqslant \hat{\sigma} \leqslant \frac{(1+\varepsilon)\gamma}{1-\varepsilon}\sigma + \delta.$$

Second, we need not memorise all the uniform variables u_{ij}'s; it suffices to remember the w_i's. When we run function sum(), we generate *fresh* variables u_{ij} conditioned on $\min\{u_{i1},\ldots,u_{i\ell}\} = d_i$. If we treat the processes errors() and sum() as a whole, then $u_{i1},\ldots,u_{i\ell}$ are indeed independent uniform random variables over $(0,1)$, so the proof gets through.

Finally, the randomness in sum() can be removed altogether. The idea is to redefine $s_i := \mathbb{E}(\sum_{j\in[\ell]} S_{ij} \,|\, d_i)$. Apparently $\mathbb{E}(s_i) = \mathbb{E}(\sum_{j=1}^{\ell} S_{ij})$; it is also well known that conditioning reduces variance: $\mathrm{Var}(s_i) \leqslant \mathrm{Var}(\sum_{j=1}^{\ell} S_{ij})$. So $\hat{\sigma}$ should still concentrate around σ. Unfortunately, it is impossible to compute s_i in the first place. But luckily, one can approximate it from two sides. Note that

$$\mathbb{E}\left(\sum_{j\in[\ell]} X_{ij} \,\middle|\, d_i\right) = \begin{cases} 0 & \text{if } d_i > \frac{\varepsilon a_i}{1+\varepsilon}, \\ 1 + (\ell-1)\frac{\varepsilon a_i/(1+\varepsilon) - d_i}{1-d_i} & \text{otherwise,} \end{cases}$$

and

$$\mathbb{E}\left(\sum_{j\in[\ell]} Y_{ij} \,\middle|\, d_i\right) = \begin{cases} 0 & \text{if } d_i > \frac{\varepsilon a_i}{1-\varepsilon}, \\ 1 + (\ell-1)\frac{\varepsilon a_i/(1-\varepsilon) - d_i}{1-d_i} & \text{otherwise.} \end{cases}$$

Hence, if we set instead

$$s_i := \begin{cases} 0 & \text{if } d_i > \varepsilon \hat{a}_i, \\ 1 + (\ell-1)\frac{\varepsilon \hat{a}_i - d_i}{1-d_i} & \text{otherwise,} \end{cases}$$

then $\hat{\sigma}$ is sandwiched between two values concentrated around σ.

Application. Let us showcase the precision sampling lemma in the context of p-th moment estimation ($p > 2$). The task is to maintain a vector $\boldsymbol{x} \in \mathbb{R}^n$ that initialises to $\mathbf{0}$ within $o(n)$ space. We support the update operation add(i, Δ) which adds value $\Delta \in \mathbb{R}$ to entry x_i. In the end we should output a multiplicative approximation to $\|\boldsymbol{x}\|_p^p := \sum_{i=1}^{n} |x_i|^p$.

Fix parameter $\varepsilon > 0$ that controls the output quality. One could estimate $\|\boldsymbol{x}\|_2$ within $(1 \pm {}^1\!/_p)$ multiplicative error using $\operatorname{polylog}(n)$ space via an algorithm by Alon, Matias and Szegedy. So after proper scaling we may assume $\|\boldsymbol{x}\|_2 \in [1 - {}^1\!/_p, 1]$, in particular $a_i := |x_i|^p \in [0, 1]$ for all $i \in [n]$. Furthermore, it follows from Hölder's inequality that $\|\boldsymbol{x}\|_p^p \geqslant \|\boldsymbol{x}\|_2^p / n^{p/2-1} \geqslant \frac{1}{\mathrm{e}\, n^{p/2-1}}$. Hence, by choosing $\delta := \frac{\varepsilon}{\mathrm{e}\, n^{p/2-1}}$, any *additive* $\pm\delta$ error to $\|\boldsymbol{x}\|_p^p$ transfers to at most a *multiplicative* $1 \pm \varepsilon$ error.

Here we lay the plan. First we play Bob's strategy to generate the deviations $d_1, \ldots, d_n$. Then we pretend to be Alice and manage the updates. In the end we as Alice must produce noisy observations $\hat{a}_i$ that deviate from truths a_i by at most multiplicative $1 \pm \varepsilon$ and additive $\pm d_i$. Finally we switch back to Bob and output an estimate $\hat{\sigma}$ to $\sum_{i=1}^n a_i = \|\boldsymbol{x}\|_p^p$. The precision sampling lemma guarantees $\frac{(1-\varepsilon)^2}{1+\varepsilon} \|\boldsymbol{x}\|_p^p \leqslant \hat{\sigma} \leqslant \frac{(1+\varepsilon)^2}{1-\varepsilon} \|\boldsymbol{x}\|_p^p$ with decent probability.

It remains to specify how to play Alice's role. Let us generate a pairwise independent function $h : [n] \to [m]$ that maps indices to bins. Independent of h, we sample another pairwise independent function $\operatorname{sgn} : [n] \to \{+1, -1\}$. Create variables $V_1, \ldots, V_m$ that are initially zero. Upon update $\operatorname{add}(i, \Delta)$ we increase $V_{h(i)}$ by $\frac{\operatorname{sgn}(i)\,\Delta}{d_i^{1/p}}$. After finishing all the updates, we would have

$$V_b := \sum_{i \in h^{-1}(b)} \frac{\operatorname{sgn}(i)\, x_i}{d_i^{1/p}}$$

for each bin $b \in [m]$. We return

$$\hat{a}_i := d_i\, |V_{h(i)}|^p$$

for each $i \in [n]$ as noisy observations.

Lemma. Take $m := (4p\,\varepsilon^{1/p-1}\,\ell^{1/p})^2$. Then for any fixed $i \in [n]$ we have $\hat{a}_i = (1 \pm \varepsilon)\, a_i \pm d_i$ with probability at least ${}^3\!/_4$.

Proof. To gain some intuition, imagine that no collision occurs in bin $h(i)$. Then $\hat{a}_i = d_i \left| \frac{\operatorname{sgn}(i)\, x_i}{d_i^{1/p}} \right|^p = |x_i|^p = a_i$ is an exact observation. But in reality there are plenty of collisions since $m \ll n$. If d_i is small then x_i contributes heavily to V_b, so the noise due to collision is negligible. On the other hand, if d_i is large then the noise might drown x_i, but it does not matter since we tolerate a large additive deviation.

Formally, let us expand the definition of $\hat{a}_i$:

$$\hat{a}_i = d_i \left| \sum_{j \in h^{-1}(b)} \frac{\operatorname{sgn}(j)\, x_j}{d_j^{1/p}} \right|^p = \left| \operatorname{sgn}(i)\, x_i + d_i^{1/p} \sum_{j \neq i : h(j) = h(i)} \frac{\operatorname{sgn}(j)\, x_j}{d_j^{1/p}} \right|^p.$$

Denote the big sum by Σ, then

$$\hat{a}_i = (|x_i| \pm d_i^{1/p} |\Sigma|)^p. \tag{1}$$

Towards showing $\hat{a}_i \approx a_i$, let us bound the variance of noise Σ:

$$\begin{aligned}
\mathbb{E}(\Sigma^2) &= \mathbb{E}\left[\left(\sum_{j \neq i} \mathbf{1}\{h(j) = h(i)\} \cdot \frac{\operatorname{sgn}(j)\, x_j}{d_j^{1/p}} \right)^2 \right] \\
&= \sum_{j \neq i} \mathbb{E}\left[\mathbf{1}\{h(j) = h(i)\} \cdot \frac{x_j^2}{d_j^{2/p}} \right] \\
&= \sum_{j \neq i} \frac{x_j^2}{m} \mathbb{E}\big(d_j^{-2/p}\big) \\
&\leqslant \frac{\ell^{2/p}}{m} = \frac{1}{(4 p\, \varepsilon^{1/p-1})^2}
\end{aligned}$$

Here in the second line, we expanded the square and noticed that a pair $\{j, j'\}$ contributes zero in expectation if $j \neq j'$. The fourth line follows from the precision sampling lemma and $\|\boldsymbol{x}\|_2 \leqslant 1$. With the variance controlled, we apply Chebyshev to obtain

$$\mathbb{P}\left(|\Sigma| > \frac{1}{2 p\, \varepsilon^{1/p-1}} \right) \leqslant \frac{1}{4}.$$

Now we assume $|\Sigma| \leqslant \frac{1}{2p\,\varepsilon^{1/p-1}} \ll 1$ and get back to (1). Let us split two cases:

(i) $d_i^{1/p} |\Sigma| \leqslant \frac{\varepsilon}{2p} |x_i|$, so the noise-signal ratio is low. We derive $\hat{a}_i = \left[\left(1 \pm \frac{\varepsilon}{2p} \right) |x_i| \right]^p \leqslant (1 \pm \varepsilon)\, a_i$.

(ii) $d_i^{1/p} |\Sigma| > \frac{\varepsilon}{2p} |x_i|$, so the noise-signal ratio is high. But since we assumed that the noise amplitude $|\Sigma|$ is tiny, the signal amplitude $|x_i|$ must be tiny too. Hence the additive deviation $|\hat{a}_i - a_i|$ should not be large.

To be precise, the assumptions imply $|x_i| < \frac{2p|\Sigma| d_i^{1/p}}{\varepsilon} \leqslant (d_i/\varepsilon)^{1/p}$. We bound

$$\begin{aligned}
\hat{a}_i - a_i &\leqslant (|x_i| + d_i^{1/p}|\Sigma|)^p - |x_i|^p \\
&\leqslant \left(|x_i| + \frac{d_i^{1/p}}{2p\,\varepsilon^{1/p-1}}\right)^p - |x_i|^p \\
&\leqslant \left[\left(\frac{d_i}{\varepsilon}\right)^{1/p} + \frac{d_i^{1/p}}{2p\,\varepsilon^{1/p-1}}\right]^p - \frac{d_i}{\varepsilon} \\
&\leqslant \frac{d_i}{\varepsilon}\left[\left(1 + \frac{\varepsilon}{2p}\right)^p - 1\right] \leqslant d_i.
\end{aligned}$$

Here the third line used that $z \mapsto (z+c)^p - z^p$ is monotonically increasing on $[-c, \infty)$, where $c \geqslant 0$.

For the inverse difference, we bound

$$\begin{aligned}
a_i - \hat{a}_i &\leqslant |x_i|^p - (|x_i| - d_i^{1/p}|\Sigma|)^p \\
&\leqslant (|x_i| + d_i^{1/p}|\Sigma|)^p - |x_i|^p \\
&\leqslant d_i
\end{aligned}$$

where the second line used convexity of $z \mapsto (z+c)^p$, and the third line follows from previous calculations. □

Of course, the success probability $3/4$ is not enough for union bound over all indices $i \in [n]$. But we can easily boost the probability to $1 - 1/n^2$ by taking the median result of $\Theta(\log n)$ independent threads. This concludes the description and analysis for Alice.